UNIVERSITE MONTPELLIER II
SCIENCES ET TECHNIQUES DU LANGUEDOC

THESE

pour obtenir le grade de

DOCTEUR DE L'UNIVERSITE MONTPELLIER II

Ecole Doctorale : **SIBAGHE**
Discipline **: Biologie de l'Evolution et Ecologie**

présentée et soutenue publiquement

par

Philippe GAYRAL

le 12 novembre 2008

Evolution des pararétrovirus endogènes de plantes : Le cas des séquences intégrées du *Banana streak virus* chez le bananier (*Musa* sp.)

JURY :

Président : Mr Jean-Loup Nottéghem (Professeur – Montpellier-SupAgro)
Rapporteur : Mr Jean-Michel Drezen (Chargé de recherche – CNRS UMR IRBI - Tours)
Rapporteur : Mme Marie-Angèle Grandbastien (Directeur de recherche - IJPB - Versailles)
Examinateur : Mme Katja Richert-Pöggeler (Directeur de recherche - Julius Kühn-Institut (JKI) Federal Research Centre for Cultivated Plants Braunschweig – Allemagne)
Co-directrice de Thèse : Mme Marie-Line Caruana (Chercheur – UMR BGPI CIRAD – Montpellier)

RESUME

Le génome des bananiers (*Musa* sp.) contient de nombreuses séquences virales EPRV (Endogenous pararetrovirus) appartenant au *Banana streak virus* (BSV), bien qu'aucun virus de plante n'ait d'étapes d'intégration dans son cycle. Certains EPRV provenant du bananier *M. balbisiana* sont infectieux car ils peuvent restituer des particules virales pathogènes en conditions de stress. La première partie de ce travail se focalise sur la biologie des EPRV. Nous avons tout d'abord analysé les caractéristiques moléculaires et génétiques des EPRV infectieux de l'espèce goldfinger du BSV (BSGFV) présents chez le bananier sauvage diploïde *M. balbisiana* cv. PKW. Nous avons ensuite identifié l'allèle infectieux de l'EPRV BSGFV, et abordé les mécanismes moléculaires de son activation par recombinaison homologue. L'évolution des séquences intégrées a été étudiée dans une deuxième partie. Une analyse phylogénétique à large échelle et une comparaison de l'évolution moléculaire des virus libres et EPRV chez trois espèces de bananiers nous ont permis de préciser l'origine phylogénétique des EPRV et de montrer que 27 évènements d'intégration indépendants se sont produits récemment dans les espèces hôtes. Nous avons ensuite étudié l'histoire évolutive de deux EPRV infectieux précédemment étudiés (BSGFV et BSV espèce Imové - BSImV) par l'analyse de leur polymorphisme de structure et de leur distribution au sein du genre *Musa*. Les résultats sont analysés en relation avec la phylogénie moléculaire des bananiers construite dans cette thèse. La probabilité d'intégration de chaque espèce de BSV est très faible, et à la différence d'autres pathosystèmes possédant des EPRV, il n'y a pas de colonisation des génomes hôtes par duplication des séquences virales une fois celles-ci intégrées. La forte diversité des EPRV chez le bananier s'explique plutôt par des évènements d'intégration indépendants de chacune des nombreuses espèces de virus libre.

Mots-clés :

Badnavirus, Bananier (*Musa* sp.), *Banana streak virus* (BSV), EPRV infectieux, Séquences pararétrovirales intégrées (EPRV).

Directeur de thèse :
Philippe Rott
Co-directeur de thèse :
Marie-Line Caruana

Discipline :
Biologie de l'Evolution et Ecologie

Unité où la thèse a été préparée :
UMR BGPI - Biologie et Génétique des Interactions Plante-Parasite
CIRAD Département BIOS
TA A-54 / K Campus international de Baillarguet
34398 MONTPELLIER CEDEX 5, France

Evolution of plant pararetroviruses: the case of integrated sequences of *Banana streak virus* in the banana genome (*Musa* sp.)

ABSTRACT

The genome of banana plants (*Musa* sp.) harbours multiple endogenous pararetrovirus sequences (EPRVs) related to *Banana streak virus* (BSV), although no virus of plants needs integration for its replication. Some EPRVs of *M. balbisiana* are able to release infectious viral genomes under stress conditions resulting in viral infection of the plant. In the first part of our work, we focused on the biological characteristics of such EPRV. We described the molecular and genetic characteristics of an infectious EPRV of BSV Golfinger species (BSGFV) present in the wild diploid *M. balbisiana* cv. PKW. We identified the infectious allele of BSGFV EPRV, and proposed a model based on homologous recombinaison for its activation. In the second part, we studied evolutionary patterns of two EPRVs previously studied (BSGFV and Imové BSV species – BSImV). We first inferred large-scale phylogenies and compared the evolution rate and selective pressures acting on non integrated virus and EPRV found in three *Musa* species. We determined the phylogenetic origin of EPRV sequences found in these *Musa* species and estimated that at least 27 independent integration events occurred recently in the genome of the host species. Then, we studied the evolutionary history of the two infectious EPRV previously described (BSGFV and BSImV species) by analysing their distribution and polymorphism of structure among representative banana species, in relation to the phylogeny of *Musa* genus reconstructed in this study. The probability of integration of every species of BSV is very weak; and unlike other pathosystems harboring EPRVs, there is no colonization of host genomes by duplication of the viral sequences once integrated. The strong diversity of EPRV in the *Musa* genome could be rather explained by independent integrations from each of the numerous BSV species.

Key words:

Badnavirus, Banana (*Musa* sp.), *Banana streak virus* (BSV), Endogenous pararetrovirus (EPRV), Infectious EPRVs.

Remerciements

Je tiens à remercier Marie-Line Caruana qui a encadré ce travail, qui a toujours été très disponible, et de bon conseil.

Mes remerciements s'adressent également à Jean Michel Drezen et Marie-Angèle Grandbastien qui ont accepté de juger le travail présenté dans ce manuscrit, ainsi qu'à Katja Richert-Pöggeler et Jean-Loup Nottéghem qui ont accepté de faire partie du jury de cette thèse.

Je remercie ensuite Marc Sitbon, Franck-Christophe Baurens, Françoise Carreel, Jean-Loup Nottéghem, Fabien Halkett et Philippe Rott qui ont eu la gentillesse de participer aux différentes réunions de comité de thèse.

Un grand merci aux membres de l'équipe 1 de l'UMR BGPI : Nathalie Laboureau, Serge Galzi et Laurence Blondin pour leur aide et Emmanuelle Muller et Matthieu Chabannes pour les nombreux échanges scientifiques. Ma reconnaissance va ensuite à Olivier Guidolin pour sa motivation et la qualité du travail accompli pendant son stage de M2.

Merci à tous les membres de l'unité qui m'ont aiguillé ou aidé, ponctuellement ou régulièrement pendant ces trois années, comme Rémi Habas pour son aide en culture in vitro. Mais surtout, un **grand** merci à Monique Royer pour ses conseils, sa disponibilité, son investissement et ses idées florissantes et quotidiennes.

Je remercie également Stéphanie Sidibe-Bocs de l'UMR Développement et Amélioration des Plantes, ainsi qu'Isabelle Hippolyte et Xavier Périer de l'UR Amélioration génétique d'espèces à multiplication végétative pour leur disponibilité et leur aide précieuse.
Je remercie Nicolas Galtier, Didier Tharreau, Gaël Thébaut, Eric Bazin et Elisabeth Fournier pour leurs relectures pertinentes.
Je tiens à remercier Dominique Lagrenée pour sa rapidité et son efficacité légendaires (mais vraies).
Je remercie le CIRAD et la région Languedoc-Roussillon pour leur soutien financier.

Je souhaite également remercier tous les non-permanents de BGPI, CBGP, LSTM, Forêt et Acridologie pour ces nombreuses réunions scientifiques passionnantes. Je remercie ma famille, et mes amis pour leur bonne humeur.

Enfin, un immense merci à Fanny pour son soutien et sa patience tout au long de cette thèse, mais aussi durant ces sept dernières années. Pour finir, mes pensées vont à Lucile, je lui dédie ce travail.

TABLE DES MATIERES

Liste des figures

Abréviations et acronymes

ADN :	Acide Désoxyribonucléique
AFLP :	Amplified Fragment Length Polymophism
ARN :	Acide Ribonucléique
ARNi :	ARN Interférent
ARNm :	ARN Messager
BAC :	Bacterial Artificial Chromosom
BEL :	BSV Expressed Locus
BSD :	Banana Streak Disease
CEO :	Capsid-encoding Organism
CIV :	Culture *In Vitro*
CRISPR :	Clustered Regularly Interspaced Short Palindromic Repeats
d_N/dS :	Taux De Substitutions Non-synonymes / Synonymes
EPRV :	Endogenous Pararetrovirus
ePVCV :	Endogenous PVCV
ERTBV :	Endogenous RTBV
ERV :	Endogenous Retrovirus
FISH :	Fluorescent *In Situ* Hybridization
GRD :	*Geminivirus*-related DNA
GVCP :	Geminiviral Coat Protein
HERV-K :	Human Endogenous Retrovirus K
IC-PCR :	Immunocapture Polymerase Chain Reaction
ICTV:	International Committee On Taxonomy Of Viruses
ISEM :	Immunosorbent Electromicroscopy
LTR :	Long Terminal Repeat
*Lyc*EPRV :	EPRV Chez *Solanum lycopersicum*
NHEJ :	Non-homologous End-joining
NsEPRV :	EPRV Chez *Nicotiana sylvestris*
*Nto*EPRV :	EPRV Chez *Nicotiana tomentosiformis*
ORF :	Open Reading Frame
PCR :	Polymerase Chain Reaction
PKW :	Pisang Klutuk Wulung
PRV :	Pararetrovirus
PTGS :	Post-transcriptional Gene Silencing
RDF :	Recombination Directionality Factors
REO:	Ribosome-encoding Organism
RGH :	Homologues De Gène De Résistance
RNaseH :	Ribonuclease H
RT :	Reverse Transcriptase
siARN :	Small Interfering RNA
*Sotu*EPRV :	EPRV Chez *Solanum tuberosum*
TGS :	Transcriptional Gene Silencing
TSD :	Target-sites Duplication

Liste des virus

ASLV :	*Avian sarcoma-leukosis virus*
BSAcVNV :	*Banana streak Acuminata Vietnam virus*
BSCavV :	*Banana streak Cavendish virus*
BSGFV :	*Banana streak GF virus*
BSImV :	*Banana streak Imové virus*
BSMyV :	*Banana streak Mysore virus*
BSOLV :	*Banana streak OL virus*
BSV :	*Banana streak virus*
CaMV :	*Cauliflower mosaic virus*
CFAV :	*Cell Fusing Agent virus*
CMV :	*Cucumber mosaic virus*
ComYMV :	*Commelina yellow mottle virus*
CSA :	*Aedes albopictus cell silent agent*
CSSV :	*Cacao swollen shoot virus*
CVMV :	*Cassava vein mosaic virus*
CMBV :	*Citrus mosaic virus*
DENV :	*Dengue Virus*
DMV-D10 :	*Dahlia mosaic virus D-10*
HIV :	*Human immunodeficiency virus*
HTDV :	*Human teratocarcinoma-derived virus*
IAPV :	*Israeli acute paralysis virus*
KRV :	*Kamiti River virus*
KTSV :	*Kalanchoe top-spotting virus*
MuLV :	*Murine leukaemia virus*
PVCV :	*Petunia vein clearing virus*
PVY :	*Potato virus Y*
RTBV :	*Rice tungro bacilliform virus*
SCBMV :	*Sugarcane bacilliform Mor virus*
ScBV :	*Sugarcane bacilliform virus*
SCYLV :	*Sugarcane yellow leafcurl virus*
TVCV :	*Tobacco vein clearing virus*
WNV :	*West-Nile Virus*
YFV :	*Yellow Fiever Virus*

INTRODUCTION

1 Les échanges génétiques ont façonné l'évolution virale

1.1 Virus et biologie évolutive

Les virus sont les entités biologiques **les plus abondantes** sur terre (Bergh et al., 1989) : des études de méta-génomique réalisées à partir d'échantillons séquencés au hasard révèlent que les gènes viraux constituent la plus grande fraction de la génosphère (Edwards & Rohwer, 2005). Ils participent à des **processus écologiques majeurs** ; les virus des océans influencent par exemple significativement la biomasse phytoplanctonique, et régulent ainsi les cycles biogéochimiques (pour revue voir (Suttle, 2005)). Il faut enfin souligner l'importance du parasitisme en évolution : la course aux armements et la coévolution hôte/parasite constituent un moteur majeur de l'évolution des organismes.

Malgré leur **importance en biologie**, les études portant sur la diversité, l'origine et l'évolution des virus sont relativement peu nombreuses par rapport à celles conduites pour d'autres organismes. La virologie est en effet une science jeune, et l'analyse de la variabilité phénotypique a longtemps été un thème majeur de l'étude des virus. Aujourd'hui, 'seulement' 1 550 espèces virales ont été décrites et classées (Astier et al., 2001), soit très peu par rapport à la biodiversité virale supposée (Edwards & Rohwer, 2005; Weinbauer & Rassoulzadegan, 2004). Durant les dernières décades, les avancées du séquençage et l'utilisation de phylogénies moléculaires ont fait avancer les connaissances sur l'évolution des virus, mais leur origine reste encore sujette à controverses. En effet, l'arbre universel du vivant est basé principalement sur des séquences ribosomales ; les virus en sont souvent exclus car ils ne possèdent justement pas de ribosomes.

1.2 Définitions des virus

Sous la forme endocellulaire, les virus sont des **éléments génétiques** capables de répliquer leur acide nucléique (une ou plusieurs molécules, ADN ou ARN, simple ou double brin, circulaire ou linéaire) indépendamment de celui de l'hôte, et de synthétiser leurs protéines. Ce sont des **parasites obligatoires** n'ayant aucun

métabolisme propre, qui utilisent les structures de la cellule hôte et notamment des ribosomes pour leur réplication, et leur assemblage en **particule virale**. Ces particules sont constituées d'acide(s) nucléique(s) porteur(s) d'information génétique, protégé(s) par une structure protéique : la capside parfois entourée d'une membrane lipidique : l'enveloppe. Les particules virales constituent le moyen utilisé par les virus pour se transmettre d'une cellule à l'autre et infecter un organisme entier, puis d'un organisme à l'autre et infecter d'autres individus de la population d'hôtes. En dehors de la cellule hôte, les particules virales sont inertes et incapables de se multiplier.

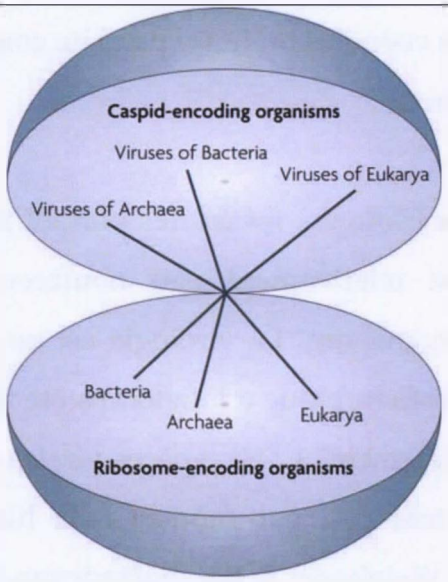

Figure 1 : Redéfinition des virus.

Représentation de l'origine des virus et des trois domaines du vivant ayant évolué depuis LUCA (last universal common ancestor). Les trois domaines du vivant possèdent des ribosomes mais n'ont pas de gènes de capside. Les virus ainsi nouvellement définis possèdent une capside, mais n'ont pas de ribosomes. Adapté de Raoult & Forterre, (2008).

Il n'existe **pas de définition universellement acceptée** des virus. L'ICTV (International Committee on Taxonomy of Viruses) les décrit comme « un système biologique élémentaire qui possède **certaines propriétés des systèmes vivants** comme le fait d'avoir un génome, ou d'être capables de s'adapter à un environnement changeant. Cependant, les virus ne peuvent pas acquérir ou stocker de l'énergie, et ne sont pas actifs en dehors de la cellule hôte ». Certains auteurs incluent les virus dans l'ensemble des entités vivantes, et les définissent comme « **un**

organisme codant une capside (capsid-encoding organism, CEO), qui s'auto-assemble dans une nucléocapside, et qui utilise un organisme codant un ribosome (ribosome-encoding organism, REO - par opposition à CEO) pour achever son cycle de vie » (Raoult & Forterre, 2008) (Figure 1).

Il existe une très forte diversité virale, et il est très probable que les virus connus à ce jour ont des **origines différentes** et qu'ils ne dérivent pas d'un même ancêtre commun, ce qui peut expliquer la difficulté d'une définition universelle des virus.

1.3 Origines des virus

Principalement à cause de l'absence de registre fossile, l'origine des virus demeure toujours inconnue. Plusieurs hypothèses non exclusives ont été formulées (Forterre, 2006; Koonin & Dolja, 2006). La première théorie explique que les virus et les cellules ont pu apparaître dans la soupe primordiale en même temps, et évoluer parallèlement (Figure 2). Dans ce scénario, au début de l'apparition de la vie, les plus anciens systèmes génétiques d'auto-réplication (probablement de l'ARN) sont devenus plus complexes et se sont enveloppés dans un sac lipidique pour aboutir au progénote à l'origine des cellules. Une autre forme réplicative parasitant les autres organismes aurait pu **garder sa simplicité** pour former des particules virales (Claverie, 2006; Koonin & Martin, 2005).

Selon la deuxième thèorie, les virus pourraient dériver de **cellules ayant subi une régression**. D'après cette hypothèse, les ancêtres des virus auraient été des êtres vivants libres ou des micro-organismes devenus des prédateurs ou des parasites de leur hôte. Il a été montré chez de nombreux taxons que le parasitisme peut entraîner la perte de nombreux gènes, dont les gènes du métabolisme apportés par l'hôte (Keeling & Slamovits, 2005; Sakharkar *et al.*, 2004; Wickett *et al.*, 2008). Ces paléo-virus auraient co-évolués avec la cellule hôte, et n'auraient conservé que la capacité à répliquer leur acide nucléique et le transfert de cellule à cellule (Suzan-Monti et al., 2006). Cette hypothèse a été récemment confortée par la découverte de virus géants au génome très complexe comme les mimivirus. Ces virus mesurent 400 nanomètres

de diamètre, et possèdent un génome de 1,2 M bases avec 1260 gènes, soit un génome deux fois plus grand que le plus grand génome viral jusqu'alors identifié, et d'une taille comparable à celui de certaines bactéries. Il apparaît qu'une trentaine de ces gènes, comme ceux codant des protéines de réparation de l'ADN ou de la traduction de l'ARN en protéines, n'ont été identifiés que dans des organismes cellulaires (Raoult et al., 2004).

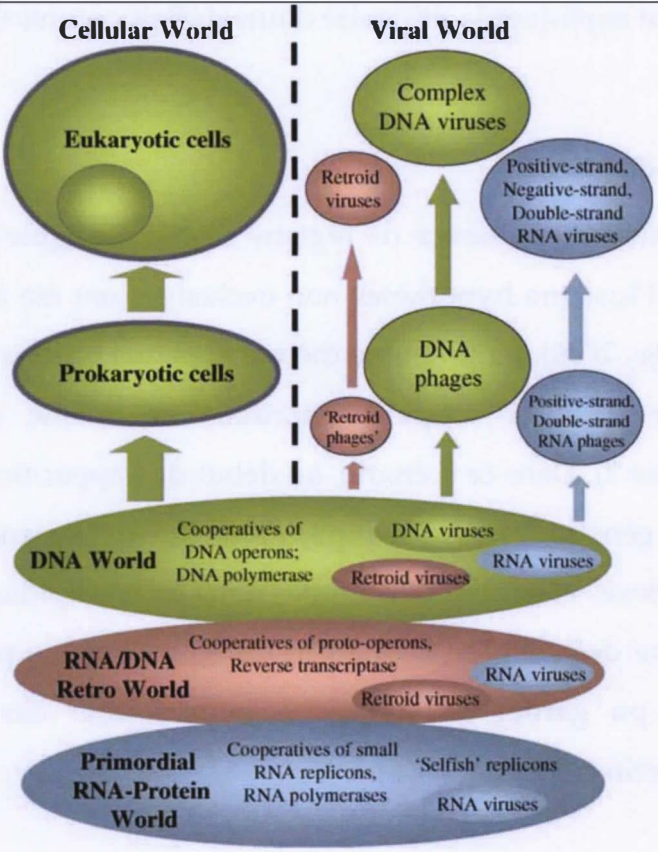

Figure 2 : Origine et évolution des organismes cellulaires et viraux (hypothèse d'une origine précoce des virus)

La direction du temps est de bas en haut. Les trois phases majeures de l'évolution pré-cellulaire sont indiquées en bas du schéma. Initialement, les différents éléments génétiques coexistaient et pouvaient se mélanger. Certains réplicons étant 'égoistes' (proto-virus), d'autres étaient 'altruistes' (proto-cellulaires). L'étape suivante est illustrée dans la partie centrale avec l'apparition des phages, qui dominaient probablement le monde viral à cette période de l'évolution. Cette étape marque la transition entre les réplicons ancestraux et les cellules procaryotes libres et leurs parasites viraux. La partie haute du schéma illustre l'émergence des cellules eucaryotes ainsi que d'une forte diversité virale. Les flèches verticales dans la partie virale du schéma indiquent les directions majeures des échanges génétiques, plutôt qu'un lien de parenté. La ligne horizontale en pointillé entre les deux 'mondes' représente les transferts génétiques horizontaux entre cellules et virus. D'après Koonin & Dolja, (2006).

Enfin, selon la troisième théorie, les virus peuvent avoir pour origine des fragments d'acides nucléiques qui se sont **« échappés » du génome cellulaire** (proto-cellules à ARN, dans la période pré-LUCA), pour devenir indépendants (Temin, 1980). Encore récemment dans l'évolution, plusieurs cas attestent de l'**émergence de rétrovirus** d'invertébrés et de vertébrés à partir d'éléments transposables dans un processus de « **capture d'enveloppe** » (Kim et al., 2004). Les rétrotransposons viraux sont des séquences d'ADN capables de se déplacer et de se multiplier au sein des génomes. Ils contiennent les ORFs *gag-pol*, sont flanqués de séquences 'long terminal repeat' (LTR) et se distinguent des rétrovirus par le fait qu'ils ne possèdent pas de gènes *env* codant une enveloppe, ce qui leur interdit une transmission de cellule à cellule, ou d'hôte à hôte. Un des exemples le plus documenté d'une émergence *de novo* de rétrovirus d'invertébrés est celui du rétrotransposon *Gypsy* de *Drosophila melanogaster* (Song et al., 1994; Terzian et al., 2001). Cet élément est devenu rétrovirus en partie après la capture d'un gène *env*-like d'un autre virus (baculovirus) (Malik *et al.*, 2000; Pearson & Rohrmann, 2002).

1.4 Pourquoi étudier les EPRV ?

Comme nous venons de le voir, l'étude des virus a très fréquemment révélé l'existence d'**échanges de matériel génétique** entre les virus et leurs hôtes, probablement facilités par la proximité physique entre les génomes lors de l'infection. Une conséquence est la présence chez les hôtes de matériel génétique fossile d'origine virale, témoin d'infections passées. L'accès aux génomes a permis de préciser l'ampleur de ce phénomène. Les intégrations virales concernent de nombreux genres viraux et se rencontrent chez de très nombreux hôtes (Leitch, 2007).

Notre travail s'intéresse aux séquences intégrées des espèces de *Banana streak virus* (BSV) chez sa plante hôte, le bananier (*Musa* sp.). Ces séquences, nommées endogenous pararetrovirus (EPRV), sont des intégrations de virus appartenant à la famille des *Caulimoviridae* infectant les plantes. Ces virus, bien que n'ayant **pas d'étapes d'intégration** dans le génome de l'hôte, y sont fréquemment retrouvés

intégrés. La particularité majeure des EPRV est la capacité, pour certains d'entre eux, à être **infectieux**, comme dans le cas des EPRV des BSV.

Ces EPRV infectieux, sont capables d'initier un cycle viral et une production de virions (*i.e.* des virus libres ou épisomaux), capables ensuite d'être **transmis** comme le virus homologue. Ces derniers participent au maintien de la diversité virale. Les EPRV sont ainsi à l'origine de l'apparition de nouvelles épidémies, et dans le contexte actuel, de la ré-émergence de la maladie de la mosaique en tiret dans les zones de cultures du bananier (Fargette et al., 2006).

L'étude des intégrations virales, et des EPRV en particulier, fournit des exemples de stratégies de réplication et de transmission des virus, de création de nouveaux virus, et permet en outre de comprendre les **conséquences évolutives** des échanges génétiques entre hôtes et parasites, un phénomène étroitement lié à l'origine et à l'évolution des virus.

2 Diversité des séquences virales endogènes

Les infections virales passées subies par une espèce ou un groupe phylogénétique peuvent laisser une trace dans le génome des organismes sous la forme de séquences virales intégrées. Les nombreuses intégrations virales observées aujourd'hui sont le reflet de la biodiversité virale contemporaine ou fossile. Les intégrations que nous avons choisi de présenter possèdent des caractéristiques biologiques originales et des histoires évolutives variées. Nous les avons utilisées par la suite comme point de comparaison pour l'étude des EPRV (cf. partie 4), et des intégrations des BSV (cf. partie 5).

Deux classes d'intégrations sont distinguées, les intégrations faisant partie du cycle de réplication viral, et les séquences intégrées de manière « accidentelles » dans le génome de leurs hôtes.

2.1 Intégrations comme stratégie de réplication virale

De nombreux virus ont évolué vers une stratégie d'intégration dans le génome de l'hôte lors de leur cycle d'infection. Cette proximité permet souvent au virus de déjouer ou détourner les défenses de l'hôte, et constitue un cas extrême de parasitisme, puisque le virus et l'hôte ne font plus qu'un. Au cours de l'évolution, ces intégrations ont fréquemment été fixées dans le génome de l'hôte et sont devenues pour certaines une composante à part entière du génome de l'espèce hôte.

Nous décrirons dans cette partie trois exemples d'intégrations virales issues du cycle de réplication des virus : les *Polydnavirus*, les phages, et les rétrovirus ; correspondant chacun à trois stratégies distinctes.

2.1.1 Le cas des *Polydnavirus*

Les *Polydnavirus* constituent probablement le cas d'intégration virale le plus extrême puisque les génomes des *Ichnovirus* et des *Bracovirus*, les deux genres de la famille des *Polydnaviridae*, n'existent que sous forme intégrée dans le génome de leurs hôtes hyménoptères (Drezen *et al.*, 2003; Dupuy *et al.*, 2006; Turnbull & Webb, 2002).

Les polydnavirus sont des symbiotes nécessaires au succès reproducteur des hyménoptères parasitoïdes du genre *Braconidae* et *Ichneumonidae*. Ils se répliquent uniquement dans les ovaires des guêpes, et sont injectés dans leurs proies : des larves de lépidoptères. Des facteurs viraux dérèglent alors leur système immunitaire (Schmidt et al., 2001) permettant aux larves de guêpes de se développer dans les chenilles. Le génome viral, physiquement lié au génome de l'hôte de manière covalente, en est totalement indissociable (Espagne et al., 2004). Une hypothèse suggère que ces virus dérivent d'un système de transfert génétique créé par les guêpes et utile au parasitoïsme. Des gènes de guêpe auraient été empaquetés dans une capside virale, elle-même « capturée » chez d'autres virus (probablement chez des *Ascovirus* et *Baculovirus*) (Dupuy et al., 2006). Les *Polydnavirus* sont des entités à la limite du monde viral : ce ne sont pas des virus autonomes, mais plutôt une production de la guêpe. Certains auteurs parlent donc des *Polydnavirus* comme d'une **sécrétion génétique** d'une organelle de guêpes (Federici & Bigot, 2003).

Pour être sécrétés, les génomes viraux seraient libérés du génome des guêpes par **excision-circularisation**. Ce mécanisme aurait lieu par **recombinaison** entre de courtes séquences répétées (une à plusieurs dizaines de nucléotides) disposés dans la même orientation, et qui encadrent les segments viraux intégrés (Annaheim & Lanzrein, 2007; Gruber *et al.*, 1996; Rattanadechakul & Webb, 2003; Savary *et al.*, 1997; Wyder *et al.*, 2002).

2.1.2 Les 'temperate phage'

Certains virus des bactéries (phages) appelés « temperate phage », ont un cycle de vie alterné. Une fraction des cellules infectées entre en phase lytique et conduit à la lyse de la cellule hôte et la libération de virions néoformés. D'autres cellules survivent à l'infection et abritent de manière permanente le génome du phage intégré à celui de la bactérie sous une forme quiescente appelée **prophage** (Campbell, 2003). Malgré de fortes contraintes sur la taille des génomes bactériens, 10 % des ORF en moyenne n'ont aucun homologue bactérien connu (ORFans) et dérivent en majorité de gènes de phages (Daubin & Ochman, 2004).

Les bactériophages sont responsables du **transfert** de nouvelles fonctions aux cellules bactériennes, et constituent un agent important de l'évolution des bactéries (Campbell, 2003; Tinsley et al., 2006). De nombreux exemples montrent que des **facteurs de pathogénicité** bactériens (comme certaines toxines) proviennent en réalité de l'expression d'ORF issus des prophages. L'interaction bactérie-prophage peut être de type **symbiotique,** lorsque le prophage est complet et fonctionnel (Wagner & Waldor, 2002). A plus long terme, les bactéries peuvent également '**domestiquer**' les prophages en utilisant les fragments viraux qui leur sont nécessaires (Banks et al., 2002). Enfin, il a été montré que les séquences de type 'Clustered regularly interspaced short palindromic repeats' (CRISPR) présentes chez la plupart des bactéries et *archaea*, dérivent des prophages acquis lors d'infections précédentes et sont fortement impliquées dans le phénotype de **résistance aux phages** qui possèdent des séquences homologues (Barrangou et al., 2007).

2.1.3 Les rétrovirus de vertébrés

Les rétrovirus ont un génome diploïde, ARN simple brin de 7 à 10 Kb, protégé par une capside et une enveloppe glycoprotéique. Une fois entré dans la cellule hôte, l'ARN génomique est synthétisé en ADN double brin par la réverse transcriptase apportée par le virus. Cet ADN rentre ensuite dans le noyau, puis s'intègre au génome de l'hôte sous forme de **provirus**, à l'aide d'une protéine virale : l'intégrase. Une fois intégré dans l'ADN de la cellule, le provirus est **stable** et se réplique avec l'ADN de l'hôte. Le provirus est alors perçu comme un gène endogène. Des ARN génomiques et des ARNm viraux sont transcrits et ces derniers sont traduits par la machinerie cellulaire.

Selon le cycle que nous venons de décrire, les rétrovirus peuvent se propager comme des agents infectieux, mais également comme des gènes cellulaires. Lorsque l'intégration se produit dans une **cellule germinale**, le provirus peut être hérité : on parle alors d'**endogénisation** (Weiss, 2006). Les rétrovirus endogènes (ERV, pour endogenous retrovirus) sont actuellement **très nombreux** dans le génome des animaux (Bromham, 2002; Griffiths, 2001). A titre d'exemple, le génome humain comporte plus de 98 000 ERV, ce qui représente environ 5 % du génome total (Lander et al., 2001; Paces et al., 2002). Ces intégrations sont la conséquence d'infections répétées et de l'endogénisation de provirus pouvant dater de plusieurs millions d'années (Johnson & Coffin, 1999). Le nombre actuel d'ERV par génome s'explique par des évènements répétés d'endogénisation, auxquels s'ajoute une amplification parfois massive de certaines lignées d'ERV à l'intérieur d'un même génome (Belshaw et al., 2004; Katzourakis et al., 2005).

En règle générale, les ERV ont subi des mutations délétères et ne sont **plus fonctionnels** (Lower, 1999). De rares cas d'intégrations pathogènes ont cependant été décrits, comme celle du groupe HTDV (*Human teratocarcinoma-derived virus*)/HERV-K (Human endogenous retrovirus K) qui sont surexprimés dans les cellules testiculaires cancéreuses chez l'homme, et associés à la formation de particules virales (Lower et al., 1996).

A l'instar des bactéries utilisant les prophages, les eucaryotes ont largement utilisé les rétrovirus endogènes pour leur propre fonction (Best *et al.*, 1997; de Parseval & Heidmann, 2005). Deux gènes de primates impliqués dans la formation du placenta sont par exemple formés en partie du gène *env* de l'enveloppe de rétrovirus endogènes des familles HERV-W (Blaise et al., 2003) et HERV-FRD (Blond et al., 2000). Un deuxième exemple, chez la souris, est celui des gènes murins de résistance au *Murine leukaemia virus* (MuLV) *Fv1* et *Fv4*, qui dérivent, respectivement, des gènes codant la capside (*gag*) et l'enveloppe (*env*) du MuLV. Le produit de *Fv4* est ainsi supposé se lier au récepteur du MuLV, et saturer les sites qui ne seraient dès lors plus disponibles en cas d'infection par le virus (Nethe et al., 2005).

2.2 Intégrations accidentelles

Un certain nombre de virus n'ayant pas d'étape connue d'intégration dans leur cycle de réplication sont néanmoins présents dans le génome de leur hôte. La plupart de ces intégrations sont des fragments tronqués du virus. Toutefois, certaines comportent un génome viral complet et fonctionnel et sont capables de reformer un virus infectieux.

2.2.1 Dans le génome des insectes

2.2.1.1 Les *Flavivirus* chez les moustiques

Les *Flavivirus* sont des virus enveloppés possédant un génome à ARN simple brin de polarité positive, d'environ 11 Kb (Fauquet et al., 2005). Leur génome code une polyprotéine qui est clivée en une protéine structurale (associée à la capside et à l'enveloppe) ainsi que des protéines non structurales (NS1-NS5). Les virus proches du groupe phylogénétique des *Flavivirus* comportent plus de 30 virus d'arthropodes pathogènes pour l'homme (par exemple le *Yellow Fiever Virus* (YFV), *Dengue Virus* (DENV), *West-Nile Virus* (WNV)), mais également des virus d'arthropodes sans hôtes intermédiaires, comme le *Cell Fusing Agent virus* (CFAV) et le *Kamiti River virus* (KRV).

C'est en utilisant des amorces PCR dégénérées amplifiant la région NS3 sur des ADN de cellules de moustique non-infectées, que Crochu et al., (2004) ont découvert la présence de séquences de *Flavivirus* intégrées (Crochu et al., 2004) appartenant à un virus alors inconnu. Ce virus nommé '*Aedes albopictus cell silent agent*' (CSA) est phylogénétiquement proche des CFAV et KRV. Trois locus d'intégration sont retrouvés dans le génome d'*Aedes albopictus*, et un locus dans l'espèce *Aedes aegypti*. Au total, les deux-tiers du génome viral ont pu être reconstitués à partir de ces fragments.

Chaque locus se compose de fragments de génomes viraux, dont le plus gros correspond à une ORF de 1557 aa. Un faible nombre d'intégrations s'est probablement produit, suivi de plusieurs évènements de brassage et dispersion des séquences dans le génome de l'hôte. Ces intégrations sont présentes chez deux espèces sœurs de moustiques, mais absentes du génome de l'espèce *Aedes w-albus*, l'espèce la plus proche phylogénétiquement. Etant donné que le CSA a divergé récemment des CFAV et KRV (environ 3500 ans), les séquences virales intégrées n'ont par conséquent pas pu être héritées d'un ancêtre commun aux deux espèces de moustiques, puisque celles-ci divergent depuis 35 Ma. Un évènement d'intégration dans chaque espèce semble être le scénario évolutif le plus parcimonieux pour expliquer la distribution des séquences intégrées CSA.

2.2.1.2 Les *Dicistrovirus* chez les abeilles

Les *Dicistroviridae* sont une famille de virus infectant les insectes, appartenant à la super famille des 'picorna-like'. Ils possèdent un génome ARN simple brin de polarité positive d'environ 9 kb, et contenant 2 ORF codant chacune une polyprotéine, clivée en plusieurs protéines lors de l'infection (Fauquet et al., 2005). Actuellement, L'*Israeli acute paralysis virus* (IAPV) serait responsable d'une mortalité accrue chez les abeilles (*Apis mellifera*), et pose de sérieux problèmes en apiculture (Cox-Foster et al., 2007).

Des intégrations d'IAPV ont été récemment décrites dans le génome de **certaines populations** d'abeilles (30 % des populations testées) ce qui suggère que le(s) évènement(s) d'intégrations sont **très récents** (Maori et al., 2007). Cinq locus d'intégration ont été identifiés et sont localisés dans les chromosomes LG3, LG5, LG6, LG12 et LG16 de l'abeille.

La deuxième particularité de ce modèle concerne le mécanisme d'intégration de ces virus (traité plus en détail dans la partie 3.2 de l'introduction). Les séquences intégrées de l'IAPV proviendraient d'une **recombinaison au niveau des ARN** génomiques viraux et des ARN des transcrits de l'hôte, puis d'une rétrotransposition en ADN dans les chromosomes de l'abeille (Maori et al., 2007).

Enfin, la troisième particularité des ce modèle est que les intégrations de l'IAPV conféreraient une **résistance** des abeilles à l'IAPV. Il existe en effet une corrélation entre la présence d'intégration et le phénotype de résistance des abeilles au virus (Maori et al., 2007). Ces résultats préliminaires doivent cependant être confirmés par des études complémentaires.

2.2.2 Dans le génome des plantes

2.2.2.1 Les *Potyviridae* chez la vigne

Les Potyvirus constituent à eux seuls le tiers des virus de plantes recensés à ce jour. Leur génome est un **ARN simple brin** de polarité positive d'environ 10 Kb (Fauquet et al., 2005). Des séquences intégrées provenant des *Potyviridae* ont été découvertes très récemment dans le génome de la **vigne** (*Vitis vinifera*) (Tanne & Sela, 2005). Seules certaines variétés de vigne possèdent des séquences potyvirales endogènes, ce qui suggère que l'intégration est **très récente** puisque ces cultivars sont issus de multiplication végétative et ne divergent que depuis quelques milliers d'années (Zohary & Hopf, 2000).

Les intégrations du *Potato virus Y* (PVY) dans le génome de la vigne proviendraient d'une recombinaison avec les ARN de l'hôte et d'une rétro-transcription dans le génome hôte, un mécanisme qu'ils partageraient avec les *Dicistrovirus* intégrés au génome de l'abeille (voir la partie 3.2).

2.2.2.2 Les *Geminiviridae* chez les Solanacées et Fabacées

La première description de séquences virales intégrées au génome de plante concerne les séquences géminivirales (Kenton et al., 1995). Les *geminiviridae* sont des phytovirus ayant un petit génome **ADN simple brin** de 2,5 à 3 Kb (Fauquet et al., 2005).

Le premier type de séquences intégrées, appelé **GRD** pour « Geminivirus-related DNA » est présent en plusieurs centaines de copies dans le génome du **tabac** (*Nicotiana tabacum*) (Bejarano et al., 1996). Ces séquences ont subi de nombreux évènements de duplication, délétion et réarrangements (Ashby et al., 1997). L'analyse plus fine des séquences a montré qu'elles forment deux groupes phylogénétiques distincts : GRD3 et GRD5, qui dérivent tous deux du même genre viral, les *Begomovirus*. En comparant la phylogénie des GRD, leur distribution dans le génome des hôtes et la phylogénie des hôtes, Murad et co-auteurs (2004) ont identifié **deux évènements d'intégrations** à l'origine de ces groupes. Les virus homologues aux séquences GRD5 se sont intégrés au génome de l'ancêtre commun aux trois espèces *Nicotiana kawakamii*, *N. tomentosa* et *N. tomentosiformis* et ces séquences sont aujourd'hui retrouvées intégrées sur le même chromosome homéologue 4. La lignée des GRD3 provient quant à elle d'une intégration plus récente dans *N. tomentosiformis* seulement, l'ancêtre paternel du tabac (*Nicotiana tabacum*) (Murad et al., 2004). L'analyse des pressions de sélection qu'a subi le groupe des GRD par l'estimation des taux de substitutions synonymes et non-synonymes (d_N/d_S) a montré que les deux groupes GRD3 et GRD5 ont évolué sous des contraintes sélectives différentes. Les auteurs remarquent également que même si les séquences GRD ne sont aujourd'hui **plus fonctionnelles**, elles gardent la trace d'une évolution sous contraintes sélectives. Ce résultat pourrait cependant provenir d'un biais d'échantillonnage des branches lors du calcul d_N/d_S : il se pourrait que les branches choisies pour l'estimation ne correspondent pas toutes à des séquences intégrées, mais reflètent également des virus libres inconnus à ce jour. Ce phénomène souligne l'importance d'un échantillonnage exhaustif des séquences lorsque des scénarios évolutifs sont issus des reconstructions phylogénétiques.

Un deuxième type de séquences géminivirales intégrées a été très récemment découvert par des analyses *in silico*. Il s'agit de motifs de quelques nucléotides de la protéine virale GVCP « geminiviral coat protein » des geminivirus retrouvés dans les domaines NB-ARC des homologues de gènes de résistance (RGH) chez différents ***Fabaceae*** (*Vigna mungo* et *V. radiata*) (Pal et al., 2007). Les auteurs de cette étude font l'hypothèse que ces séquences virales ont été intégrées lors d'une interaction plante-pathogène, et sont depuis maintenues par sélection purificatrice. Ces motifs viraux pourraient en effet être utilisés par les plantes dans la défense contre les *Geminiviridae*, au travers d'une reconnaissance spécifique des transcrits viraux et de leur dégradation.

2.2.2.3 Les *Caulimoviridae* chez les plantes

Les *Caulimoviridae* (ou pararétrovirus de plante) constituent la troisième famille virale intégrée dans le génome des plantes. Les intégrations de cette famille virale sont appelées EPRV pour endogenous pararetrovirus.

Parmi toutes les intégrations accidentelles, les EPRV sont les plus abondamment étudiés pour différentes raisons. Les EPRV sont tout d'abord retrouvés dans le génome d'un **grand nombre de plantes** appartenant à des familles très différentes, et un même génome peut abriter de **nombreuses** copies d'EPRV. L'intégration n'est donc pas un phénomène isolé. Ensuite, certains EPRV sont **infectieux**, car capables de restituer un génome viral complet et fonctionnel. Cette caractéristique est unique parmi les autres cas d'intégration accidentelle, et les rapproche plus des provirus ou des prophages. Enfin, les EPRV sont fortement suspectés d'avoir été **domestiqués** par les plantes, et de jouer un rôle dans la résistance acquise contre les virus correspondants.

3 Comment les virus deviennent ils endogènes ?

La plupart des virus endogènes ont un génome ADN, ce qui a probablement facilité leur intégration dans le génome ADN des hôtes. Certaines intégrations correspondent cependant à des virus au génome ARN, et nous verrons dans la partie suivante comment ils ont pu néanmoins s'intégrer. Nous aborderons dans cette partie les principaux mécanismes moléculaires décrits ou supposés, qui ont permis aux séquences virales de s'intégrer au génome de l'hôte.

3.1 Endogénisation par des enzymes spécialisées

Deux grands groupes viraux, les *Retroviridae* et certains bactériophages, ont dans leur cycle de réplication des mécanismes actifs permettant l'intégration de leur génome dans celui de la cellule hôte.

Le premier exemple est celui des bactériophages, qui utilisent des enzymes responsables de l'entrée et la sortie du génome viral. Leur intégration dans les chromosomes de l'hôte se fait par l'action **d'intégrases** qui effectuent des **recombinaisons sites-spécifiques** entre les sites nommés attP du génome des phages, et les sites attB des génomes bactériens (Nash, 1981). L'intégration génère des sites jonctions (attL et attR) qui flanquent le prophage, et qui sont utilisés pour l'**excision du génome viral** lors d'une induction. L'intégrase virale (gpInt) catalyse à la fois l'intégration (recombinaison au niveau des sites attP et attB) et l'excision (recombinaison sur les sites attL et attR). L'issue de la réaction est déterminée par des protéines accessoires (recombination directionality factors ; RDFs) (Ghosh et al., 2006).

Le deuxième exemple concerne les rétrovirus. Tous les rétrovirus ont en commun le mécanisme d'intégration réalisé par des monomères d'**intégrase** liés au génome viral (Figure 3). Chez certains rétrovirus, comme l'*Avian sarcoma-leukosis virus* (ASLV, *Alpharetrovirus*), les sites d'intégrations sont aléatoires (Mitchell et al., 2004; Narezkina et al., 2004). D'autres en revanche, ont des **sites d'intégrations préférentiels** (Lewinski et al., 2006) qui sont en général des zones actives du génome.

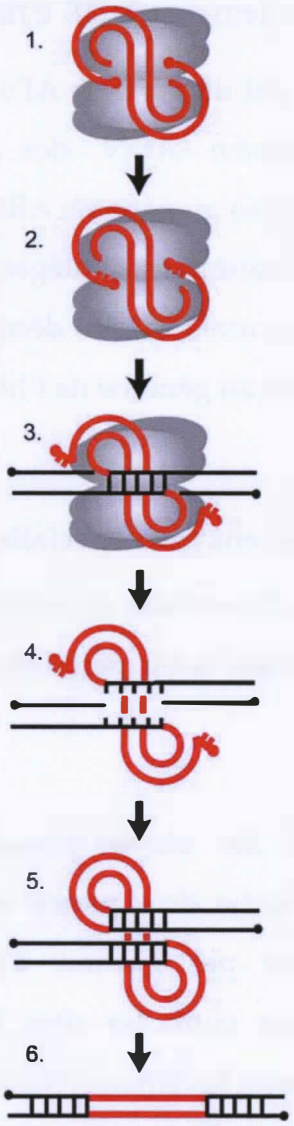

Figure 3 : Mécanisme d'intégration des rétrovirus

Les monomères d'intégrase sont représentés par les ovales gris. Les lignes rouges indiquent l'ADN de rétrovirus et les lignes noires l'ADN-cible dans le génome de l'hôte. Les points en bout de séquence représentent les extrémités 5' de la molécule d'ADN. **(1)** L'ADN viral linéaire possédant des bords francs est lié au complexe de pré-intégration. **(2)** L'intégrase retire deux nucléotides des extrémités 3' de l'ADN viral. **(3)** L'intégrase ligue les extrémités 3' de l'ADN viral à l'ADN cible, qui a préalablement subi une cassure. **(4)** L'ADN cible est déshybridé entre les deux extrémités jointes de l'ADN viral, formant des gaps dans l'ADN cible. **(5)** Les nucléotides manquants sont ajoutés par plusieurs enzymes de réparation. **(6)** Le provirus est alors flanqué par des fragments répétés de l'ADN cible (TSD – Target-sites duplication). Adapté de Lewinski et al. (2006).

Le *Human immunodeficiency virus* (HIV, *Lentivirus*) s'intègre préférentiellement dans les gènes transcrits (Mitchell et al., 2004; Schroder et al., 2002), et l'intégration du *Murine leukemia virus* (MuLV ; *Gammaretrovirus*) est biaisée vers les sites d'initiation

de la transcription de gènes (Wu et al., 2003). A la différence des bactériophages, le génome intégré - ou **provirus,** ne ressort pas du chromosome de l'hôte. Il est en effet directement **transcrit** grâce aux promoteurs viraux présents dans les séquences LTR flanquantes.

3.2 Intégration par recombinaison des ARN

En dehors des rétrovirus et des phages dont la réplication nécessite une intégration de leur génome, il existe de nombreux autres cas de présence de séquences virales dans le génome de leurs hôtes, sans que cela ne soit lié à un mécanisme connu de réplication. Un premier mécanisme est supposé expliquer la présence de séquences virales des ***Potyviridae*** dans le cas du PVY chez la vigne (Tanne & Sela, 2005) (présentées dans la partie 2.2.2.1), et des ***Dicistroviridae*** dans le cas de l'IACV chez l'abeille (Maori et al., 2007) (cf. partie 2.2.2.2). Ces deux virus possèdent des génomes ARN.

Des fragments de ces deux virus sont intégrés au génome de leur hôte respectif, mais de manière surprenante, des séquences correspondant à des ARN cellulaires ont également été retrouvées **intégrées à des génomes viraux** défectifs, preuve d'une interaction au niveau ARN. Ilan Sela et son équipe ont donc fait l'hypothèse d'une **recombinaison au niveau ARN :** les génomes viraux se seraient recombinés avec des ARN hétérologues de la cellule hôte. Le mécanisme moléculaire d'intégration reste inconnu, mais les auteurs proposent un mécanisme de **cassure et réparation** qui permettrait de former de manière aspécifique un ARN chimérique composé de deux séquences hétérologues (Gallei et al., 2004). Ces ARN chimériques seraient ensuite rétrotransposés par les mécanismes propres aux rétroéléments (Zhang, 2003) pour former les intégrations observées aujourd'hui (Maori *et al.*, 2007; Tanne & Sela, 2005).

3.3 Intégration par recombinaison non homologue

La recombinaison non-homologue (ou « illégitime ») est le mécanisme qui a été proposé pour expliquer les intégrations des ***Caulimoviridae*** (Staginnus & Richert-Pöggeler, 2006) et des ***Geminiviridae*** (Bejarano et al., 1996).

Chez les eucaryotes, les cassures double brins se produisent de manière inévitable dans le génome de la cellule hôte, et sont fréquemment réparées par recombinaison non homologue selon le mécanisme de « **non-homologous end-joining** » (NHEJ) (Puchta, 2005). Un fragment d'**ADN simple brin** non homologue s'hybride au niveau de la cassure, qui est alors réparée par synthèse d'ADN complémentaire à partir du brin non homologue.

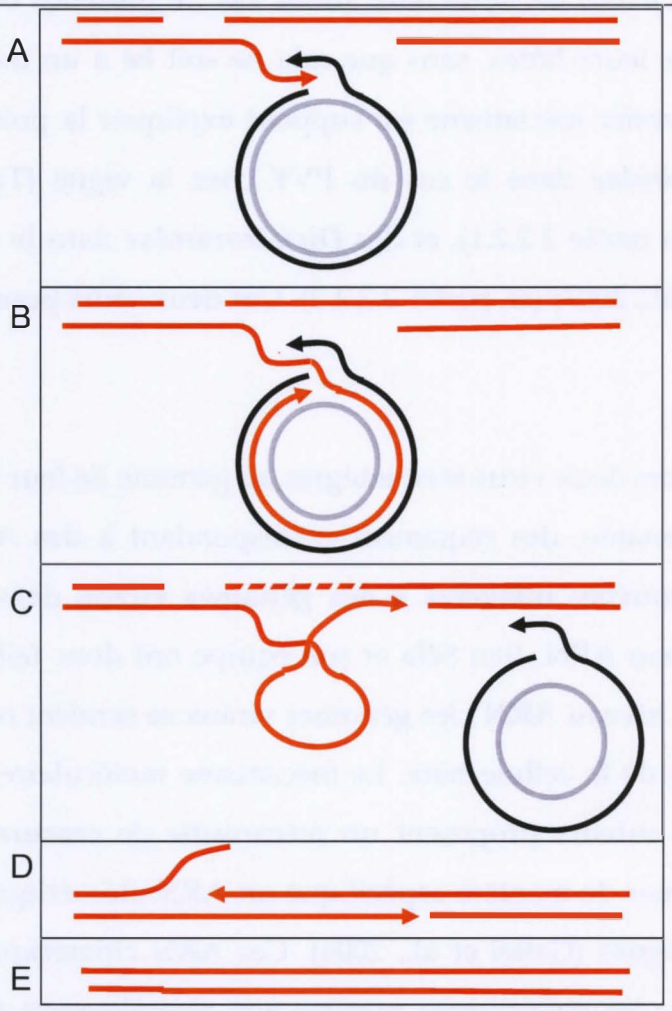

Figure 4 : Mécanisme d'intégration des EPRV et GRD par recombinaison non homologue

L'ADN viral est représenté en noir, celui de l'hôte en rouge. Les brins néosynthétisés sont indiqués en orange. **(A)** Une recombinaison non homologue peut s'initier par NHEJ (non-homologous end-joining) lorsque des fragments simples brin générés au niveau des cassures du génome hôte, s'apparient sur des micro-homologies à des séquences virales hétérologues, plutôt qu'à des séquences cellulaires homologues. Les matrices virales simple brin peuvent être soit les intermédiaires ssDNA des *Geminiviridae*, soit les courtes séquences simples brins présents au niveau des interruptions de séquences du génome des *Caulimoviridae*. **(B-D)** Les enzymes de réparation de l'ADN synthétisent les brins complémentaires à la séquence virale, réparent les nucléotides manquants et les cassures par ligation. **(E)** La séquence virale devenue endogène est maintenant intégrée de manière stable dans le génome de la cellule hôte. D'après Hohn et al. (2008).

En contexte d'infection, le génome viral entre dans le noyau des cellules ; les génomes viraux et cellulaires sont alors physiquement proches. Le NHEJ peut alors avoir lieu entre les deux génomes (Figure 4), d'autant plus facilement que les *Geminiviridae* ont un génome simple brin, et qu'il existe dans le génome des *Caulimoviridae* de courtes séquences simple brin au niveau des deux **interruptions de séquences** (Jakowitsch et al., 1999) (voir la partie 5.2.2 pour plus de détails sur le génome des *Caulimoviridae*). L'étude de la structure des EPRV a permis de confirmer que la zone présentant les interruptions de séquences constitue une **région privilégiée** pour l'intégration du ePVCV (endogenous PVCV) chez le pétunia (Richert-Pöggeler et al., 2003), du NsEPRV chez le tabac (Jakowitsch et al., 1999) et du ERTBV (endogenous *Rice tungro baciliform virus* - RTBV) chez le riz (Kunii et al., 2004).

La présence de régions non transcrites dans les **GRD** (Geminivirus-related DNA) suggère que l'intégration passe par un intermédiaire ADN, et non ARN. Ce mécanisme serait donc à l'origine d'intégrations de génomes, ou fragments de génome viral ADN de manière **aléatoire et non dirigée** dans le génome de l'hôte.

4 Les pararétrovirus endogènes de plante

4.1 Diversité des virus intégrés et des hôtes

4.1.1 EPRV non infectieux

La majorité des EPRV sont composés de séquences virales ayant subi des mutations leur faisant perdre toute capacité infectieuse. De tels EPRV sont par exemple formés d'un génome viral incomplet, avec des ORF pouvant être tronquées. Des réarrangements du génome des plantes postérieurs à l'intégration peuvent également conduire à des duplications, des insertions/délétions (indels) ou des inversions des fragments viraux qui deviennent alors non fonctionnels. Enfin, certains EPRV ont accumulé des mutations délétères lors de l'évolution du génome de l'hôte. Dans la région codante par exemple, des codons stop prématurés peuvent être introduits par des substitutions ou par les indels qui décalent la phase de lecture.

Bien que non fonctionnels, ces EPRV peuvent être importants pour les virus homologues, et pour l'évolution du génome de l'hôte (hypothèses et exemples développés dans la partie 5.3.4). Trois exemples d'EPRV non-infectieux sont présentés dans les parties suivantes.

4.1.1.1 Les ERTVB chez le riz

L'analyse de la séquence génomique complète du riz *Oryza japonica* cv. Nipponbare a révélé la présence de 29 séquences dispersées dans le génome du riz et proches du *Rice tungro bacilliform virus* (RTBV). Ces intégrations, appelées ERTBV (pour endogenous RTBV) ont pu être attribuées à trois groupes phylogénétiques qui présentent plus de 80 % d'identité nucléotidique entre eux (Kunii et al., 2004).

Les auteurs supposent que les ERTBV ne sont pas infectieux. En effet, ces séquences sont d'une part fortement réarrangées et comportent des délétions et insertions, des inversions ainsi que des duplications. D'autre part, aucun virus complet n'est présent au sein des locus d'intégration. Enfin, aucune particule virale correspondante n'a été détectée dans les plants de riz hébergeant ces séquences virales.

Les auteurs ont pu cependant reconstruire *in silico* la séquence consensus du virus circulaire à l'origine de ces multiples locus d'intégration. De manière surprenante, les virus reconstruits ne possèdent pas d'homologue de l'ORF II des RTBV, et les autres ORF présentent une forte dissimilarité nucléotidique par rapport aux ORF homologues chez le RTBV (ex. ORFI : 49 % d'identité).

Kunii et coauteurs, (2004) ont identifié des sites viraux privilégiés pour l'intégration. La région intergénique et la région des discontinuités du génome chez les autres *Caulimoviridae* sont en effet surreprésentées par rapport aux autres. Un court fragment ADN simple brin au niveau de ces discontinuités est en effet supposé servir de matrice à la recombinaison non homologue qui permet l'intégration (cf. partie 3.3) (Hohn et al., 2008). De même, les jonctions des sites d'intégrations sont dans la

majorité des cas des répétitions de dimères AT. Les auteurs suggèrent que ces séquences pourraient médier l'intégration, mais leur origine exacte : virale (queue poly-AT accrochée à la séquence virale), ou cellulaire (site d'intégration riche en poly-AT), n'est pas connue.

De nombreux ERTBV ont été retrouvés par Southern blot chez quatre espèces de riz d'origine Sud-Asiatique ou Australienne, alors que les trois espèces strictement Africaines n'ont pas, ou très peu, d'intégrations. Les auteurs ont remarqué que le nombre d'ERTBV était corrélé au degré de méthylation des séquences intégrées, mais aussi au degré de résistance au RTBV. L'étude de ce modèle n'en est qu'à son début, et les pistes concernant la régulation et l'utilisation des ERTBV par les plantes sont prometteuses.

4.1.1.2 Les EPRV proches du TVCV chez les *Solanaceae*

Ces dernières années, les travaux sur les intégrations de virus proches du *Tobacco vein clearing virus* (TVCV) dans les solanacées ont fait significativement avancer les connaissances sur le contrôle épigénétique des EPRV et sur leur utilisation par les plantes (voir la partie 4.2) (Hohn *et al.*, 2008; Staginnus & Richert-Pöggeler, 2006). Ces intégrations ont été retrouvées dans le génome de nombreuses espèces, mais les relations phylogénétiques entre virus ainsi que le nombre d'intégrations expliquant la présence de ces EPRV sont encore mal connus.

Ces EPRV possèdent plusieurs caractéristiques communes. Les intégrations sont en général **nombreuses**, dispersées dans le génome mais restreintes à l'hétérochromatine et insérées préférentiellement dans les **régions péricentromériques** non codantes des chromosomes (Hohn *et al.*, 2008; Staginnus & Richert-Pöggeler, 2006). Outre le cas des EPRV du TVCV chez les hybrides de tabac, aucun cas d'activation d'EPRV de solanacées n'a été reporté à ce jour. Ces EPRV ont généralement des ORF défectives et ne peuvent **pas être infectieux**. De plus, les virus épisomaux correspondant aux intégrations ne sont en général pas connus. Les EPRV décrits à ce jour chez les solanacées ont été classés en cinq 'types' : *Ns*EPRV,

*Nto*EPRV, *Sotu*EPRV, *Lyc*EPRV et TVCV EPRV (présentés dans les parties 4.1.1.2 et 4.1.2.1), en fonction de l'espèce de l'hôte et de la séquence nucléotidique virale.

Les ***Ns*EPRV** ont été le premier type d'EPRV décrits. Ils ont été identifiés par hybridation de banques λ de tabac *N. tabacum*, et détectés dans les autres espèces de tabac par Southern blots avec des sondes virales correspondant à la région de la réverse transcriptase. Ainsi, environ 10^3 copies de cet EPRV est présent par génome diploïde de *Nicotiana sylvestris*, et dans l'espèce *N. tabacum* (Jakowitsch et al., 1999). Cette dernière provient du croisement de *N. sylvestris* (possédant des EPRV) avec *N. tomentosiformis* dont le génome ne contient pas ces EPRV. Le séquençage de 22 clones génomiques λ de *N. tabacum* a permis, après assemblage des différents fragments défectifs, de reconstruire *in silico* le génome du virus correspondant. Ce nouveau virus ainsi identifié, inconnu à l'état libre, s'est révélé être proche du **TVCV** (découvert plus tard) avec lequel il partage 80 % d'identité nucléotidique au niveau des ORF, et 60 % dans les régions non codantes (Lockhart et al., 2000).

Les ***Nto*EPRV** forment le deuxième type d'EPRV. Ils sont associés au génome *N. tomentosiformis*, qui abrite environ 600 copies (Gregor et al., 2004). Ces EPRV sont beaucoup plus nombreux (approximativement 4.10^3 copies) dans le génome de *N. tabacum*, mais sont par contre absents du génome de *N. sylvestris*, l'autre progéniteur de *N. tabacum*. Selon les auteurs, cet écart entre le nombre de copies pourrait s'être creusé lors de l'hybridation entre les diploïdes *N. tomentosiformis* et *N. sylvestris*, ayant conduit au génome polyploïde de *N. tabacum*. Deux modèles ont été proposés: le premier est une **élimination** des nombreux *Nto*EPRV de *N. tomentosiformis* lors de l'hybridation. Le deuxième modèle suppose une **amplification** chez *N. tabacum*, à partir d'un faible nombre d'EPRV présents depuis l'hybridation (il y a 10 000 ans) chez *N. tomentosiformis* (Jakowitsch et al., 1999; Matzke et al., 2004). A nouveau, la reconstitution *in silico* du virus initial à partir des EPRV conduit à un virus phylogénétiquement proche du TVCV et des *Ns*EPRV. L'analyse des jonctions des 24 intégrations de *Nto*EPRV observées chez *N. tomentosiformis* révèle que dans $^2/_3$ des cas, des rétroéléments de type Gypsy/Ty3 sont retrouvés dans l'environnement immédiat, ou de part-et d'autre des EPRV. Les auteurs font l'hypothèse que ces

rétroéléments auraient pu jouer un rôle majeur dans (1) l'intégration des EPRV par recombinaison avec un ARN pré-génomique viral et formation d'un rétroélément chimérique pouvant poursuivre sa rétrotransposition ; (2) dans l'amplification des EPRV qui suivrait l'activation des rétroéléments ; et (3) dans l'élimination des EPRV lors d'une délétion des chimères EPRV-rétroélément par l'hôte (Matzke et al., 2004).

Le troisième type d'EPRV comprend les ***Sotu*EPRV** proches du TVCV intégrés dans la pomme de terre (*Solanum tuberosum*) (Hansen et al., 2005). Deux familles ont pu être identifiées : elles sont présentes sur 36 des 48 chromosomes et fréquemment trouvées dans les régions centromériques. Les homologues des *Sotu*EPRV ont été trouvés dans des banques EST de tomate et de pomme de terre exposées à des stress d'infection par *Agrobacterium*. Les auteurs en concluent que l'expression de ces EPRV pourrait être liée à des mécanismes épigénétiques de défense contre les gènes viraux (hypothèse développée dans la partie 4.2).

Le quatrième type appelé ***Lyc*EPRV** est trouvé chez la tomate cultivée (*S. lycopersicum*) et les espèces sauvages proches (*S. cheesmaniae* et *S. pimpinellifolium*) (Budiman et al., 2000; Staginnus et al., 2007). Deux autres espèces sauvages (*S. habrochaites* et *S. peruvianum*) présentent des profils d'hybridation nettement distincts, laissant supposer une différence dans la nature ou le nombre des EPRV (Staginnus et al., 2007). Les analyses de dissimilarité montrent que les *Lyc*EPRV sont plus proches des EPRV du tabac que des *Sotu*EPRV. Ce résultat suggère que les *Lyc*EPRV et *Sotu*EPRV, bien qu'intégrés dans des plantes appartenant au même genre, n'ont pas été hérités à partir d'un ancêtre commun, mais correspondent plutôt à deux intégrations indépendantes.

4.1.1.3 Le DMV-D10 chez les *Dahlia*

Jusqu'à présent, 4 genres viraux sur les 6 que comporte la famille *Caulimoviridae* étaient connus pour avoir des EPRV. Deux genres, *Soymovirus* et *Caulimovirus*, ne semblaient pas être concernés par ce phénomène. Ce n'est que très récemment que des EPRV d'un nouveau virus du genre *Caulimovirus*: le *Dahlia mosaic virus* D-10

(DMV-D10) (Pahalawatta et al., 2008b), ont été décrits dans le génome du dahlia (*Dahlia variabilis*) (Pahalawatta et al., 2008a).

Les travaux sur ce modèle n'en sont qu'à leurs débuts, ils renforcent néanmoins l'idée que les EPRV sont fréquents chez les plantes, et que le genre viral de l'emblématique *Cauliflower mosaic virus* (CaMV) abondamment utilisé en biologie moléculaire, est également concerné par les intégrations. Enfin, certains EPRV du DMV pourraient être infectieux et leur activation expliquerait l'apparition du virus en dehors de contaminations virales extérieures. Cette hypothèse reste néanmoins à être confirmée.

4.1.2 EPRV infectieux

Nous présenterons dans cette partie les trois cas connus de pathosystèmes possédant des EPRV infectieux. Chaque pathosystème diffère quant au genre viral intégré ainsi qu'aux plantes hôtes, ce qui conduit à une histoire évolutive qui leur est propre. Chaque cas est donc caractérisé par un nombre d'EPRV par génome, un ou plusieurs mécanismes d'activation, un ou des mécanismes de contrôle des EPRV par les plantes. *In fine*, les conséquences des EPRV pour les plantes ne seront pas les mêmes.

Malgré ces différences, plusieurs caractéristiques leurs sont communes. Tous les EPRV infectieux possèdent les informations génétiques nécessaires à la restitution d'un génome viral complet et fonctionnel. Par ailleurs, les EPRV sont intégrés fréquemment à proximité de rétroéléments. Nous utiliserons les résultats de cette thèse pour discuter de ces caractéristiques à la fin de ce document

4.1.2.1 Les EPRV TVCV chez le tabac *N. edwardsonii*

Le *Tobacco vein clearing virus* (TVCV) est un *Caulimoviridae* du genre *cavemovirus* possèdant un génome de 7,8 Kb et qui comporte quatre ORF (Figure 5). Cette organisation génomique est également partagée par l'autre espèce connue de ce genre : le *Cassava vein mosaic virus* (CVMV) (Fauquet et al., 2005). Le TVCV est retrouvé chez l'espèce hybride *Nicotiana edwardsonii* seulement, chez laquelle il

provoque des symptômes foliaires tardifs, et est transmis uniquement par les graines. Des essais de transmission mécanique et via des pucerons n'ont pas permis d'infecter sept autres espèces de solanacées (Lockhart et al., 2000).

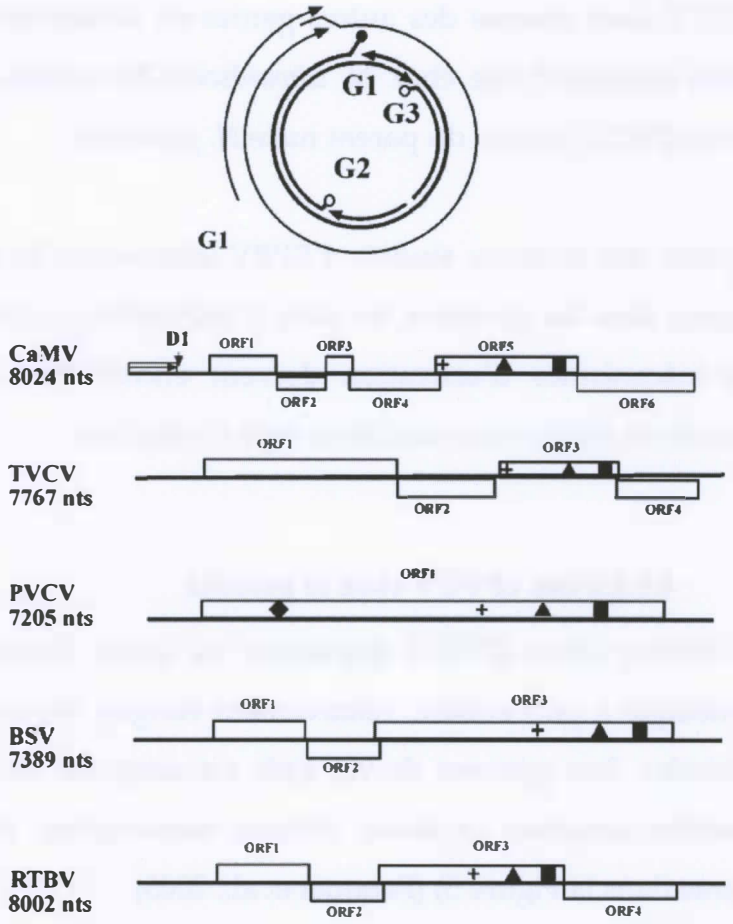

Figure 5 : Génome des *Caulimoviridae*

En haut: génome circulaire du CaMV, les transcrits sont représentés par les lignes plus fines. Le génome des caulimovirus possède trois discontinuitées (G1 à G3). Chez les badnavirus, seulement deux discontinuités sont présentes sur chacun des deux brins. En bas : Organisation génomique de virus appartenant à d'autres genres viraux des *Caulimoviridae*, dont le BSV. Les rectangles représentent les ORF. La position des motifs typiques des *Caulimoviridae* et des rétro-éléments est indiquée par les symboles : aspartate protéase (+), Reverse transcriptase (▲), Ribonucléase H (■), et intégrase (♦). D'après Harper et al. (2002).

N. edwardsonii est un hybride interspécifique hexaploïde entre les espèces *N. clevelandii* et *N. glutinosa*. Puisqu'aucun symptôme ni virus n'étaient détectés chez les parents (Jakowitsch et al., 1999; Lockhart et al., 2000), et que le TVCV apparaissait *de novo* chez les hybrides *N. edwardsonii*, Lockhart et co-auteurs (2000) ont fait

l'hypothèse que des séquences virales infectieuses étaient intégrées dans le génome de *N. edwardsonii*. Des séquences intégrées du TVCV ont été détectées par Southern blot dans le génome *N. glutinosa*, mais pas chez l'autre parent *N. clevelandii*. Une faible hybridation a également été observée avec les génomes de *N. tabacum* et *N. rustica*, probablement dû à la présence de séquences EPRV génétiquement proches. Les EPRV du TVCV sont absents des autres genres de solanacées (Lockhart et al., 2000). Les auteurs concluent que chez *N. edwardsonii* les virions proviennent des EPRV infectieux du TVCV, hérités du parent mâle *N. glutinosa*.

Ce modèle fait partie des trois cas étudiés d'EPRV infectieux. Des pistes concernant le nombre de copies dans les génomes, les sites d'intégration, la structure des EPRV ainsi que leurs mécanismes d'activation doivent encore être explorés afin de comprendre le mode de fonctionnement de ce type d'infection.

4.1.2.2 Les ePVCV chez le pétunia

Le *Petunia vein clearing virus* (PVCV) appartient au genre *Petuvirus*. Il forme un groupe phylogénétique à part entière, relativement éloigné des autres genres de la famille *Caulimoviridae*. Son génome de 7,2 Kpb est composé de deux longs ORF contenant les motifs aspartate protéase, réverse transcriptase et ribonucléase H (génome représenté dans la Figure 5) (Fauquet et al., 2005).
Cinq locus d'intégration ont été identifiés chez *P. hybrida* par FISH, dont le génome abriterait 100 à 200 copies du virus (Richert-Pöggeler et al., 2003). Ces intégrations (ou ePVCV) se composent de **répétitions en tandem** du génome complet du PVCV. L'espèce *P. hybrida* provient d'une hybridation entre les espèces sauvages *P. integrifolia* ssp. *inflata*, et *P. axillaris* ssp. *axillaris*, cette dernière seulement lui aurait transmis les EPRV.

Ce virus a créé la surprise lors de la découverte dans son génome, d'une région de l'ORFI très similaire aux domaines catalytiques typiques des **intégrases** (HHCC et $DD_{35}E$) (Richert-Pöggeler & Shepherd, 1997). Ce type de motif est également présent dans les intégrases des rétrovirus et rétrotransposons, et utilisé pour intégrer l'ADN

viral à celui de l'hôte (Figure 3). Un tel motif intégrase aurait pu aisément expliquer la présence d'EPRV du PVCV dans le génome des pétunias (Richert-Pöggeler & Shepherd, 1997). Cependant, l'hypothèse d'une intégrase fonctionnelle chez les PVCV n'est plus avancée, car aucun autre motif typique des intégrases fonctionnelles n'a été retrouvé chez le PVCV. Ce motif résulterait vraisemblablement d'une enzyme ayant été fonctionnelle dans le passé, mais qui ne serait plus utilisée par les PVCV contemporains (Harper et al., 2002; Richert-Pöggeler et al., 2003).

Richert- Pöggeler et co-auteurs (2003) ont été les premiers, et les seuls à ce jour, à apporter la preuve expérimentale de la **nature infectieuse** d'un EPRV. Les auteurs ont introduit par biolistique un clone génomique λ de *P. hybrida* correspondant à un fragment d'ePVCV, et contenant l'équivalent d'une copie du génome viral. Ce clone, bombardé sur des feuilles de *P. parodii* dont le génome est indemne d'ePVCV, a restitué un génome viral fonctionnel. Selon la construction utilisée, jusqu'à 50 % des plantes ont présenté des symptômes typiques du virus, et des particules virales de PVCV ont été observées sous 6 à 8 semaines après bombardement (Richert-Pöggeler et al., 2003).

Selon les auteurs, le mécanisme moléculaire privilégié responsable de l'activation des ePVCV serait une **transcription de l'EPRV** plutôt qu'une recombinaison. En effet, bien qu'étant linéaires, les EPRV sont composés de copies en tandem du génome viral, et possèdent donc une structure linéaire identique à la molécule circulaire du génome viral. La transcription des EPRV suffirait à produire un ARN prégénomique viral. Au cours de l'expérience de biolistique, le clone λ a donc probablement servi de vecteur d'expression transitoire de l'ePVCV qu'il contenait. Des travaux récents sur le contrôle épigénétique des ePVCV ont confirmé que les séquences intégrées sont en effet faiblement transcrites chez *P. hybrida* (Noreen et al., 2007).

4.1.2.3 Les EPRV BSV chez le bananier

Les séquences endogènes des BSV sont le troisième modèle d'EPRV connu qui possède des EPRV infectieux, et seront décrites en détail dans la partie 5.

Ce pathosystème est également l'un des premiers étudiés pour ses EPRV infectieux, car il est le seul qui touche une culture économiquement importante. Il constitue également pour des raisons qui seront développées plus loin un système biologique intéressant pour étudier les aspects plus fondamentaux liés à la diversité des intégrations, aux mécanismes d'activation et à l'évolution des EPRV infectieux.

4.2 Régulation épigénétique des EPRV

Nous avons vu précédemment que les EPRV infectieux sont capables de restituer des virions, à l'aide d'un mécanisme impliquant probablement la recombinaison au niveau de l'EPRV et/ou une transcription des séquences intégrées (hypothèses développées dans la partie 5.3.2). Comme pour les rétrotransposons, la délétion des EPRV parasites du génome n'est pas toujours possible, et les hôtes ont développé des mécanismes épigénétiques pour stopper leur invasion par rétrotransposition, notamment en régulant leur transcription (Kidwell & Lisch, 2000; Slotkin & Martienssen, 2007).

4.2.1 Les EPRV sont contrôlés par les plantes

Les études portant sur les EPRV des solanacées (genres *Petunia* et *Nicotiana*) ont amené des preuves de l'utilisation par la plante de mécanismes épigénétiques pour réprimer l'activation des EPRV (Figure 6a) [voir Staginnus et al., (2006) et Hohn et al., (2008) pour review]. Selon le modèle proposé, l'ADN des EPRV (qu'ils soient infectieux ou non), est **méthylé** (Mette et al., 2002; Noreen et al., 2007; Richert-Pöggeler et al., 2003; Staginnus et al., 2007). Du fait de l'observation des intégrations dans les régions péricentromériques des chromosomes, et des modification des **histones** (Noreen et al., 2007), ces EPRV sont principalement localisés dans l'**hétérochromatine**. Ces modifications épigénétiques sont utilisées par les plantes pour réguler l'expression de gènes endogènes, mais permettent également une protection du génome contre les dommages liés à la ré-activation des transposons. Dans le cas des EPRV, cette protection aurait pour conséquence une réduction

significative de la transcription des EPRV, voire son extinction, se traduisant par la suppression de l'activation des EPRV.

Une **faible transcription** des EPRV est néanmoins observée par RT-PCR pour les intégrations suivantes : ePVCV (Noreen et al., 2007; Richert-Pöggeler et al., 2003), *Lyc*EPRV (Staginnus et al., 2007), *Sotu*EPRV (Hansen et al., 2005) et *Ns*EPRV (Mette et al., 2002). Ces transcrits ARN peuvent alors former des structures secondaires double brin, et servir de précurseurs pour la formation de siARN (small-interfering RNA), comme observés chez le pétunia et la tomate (Noreen et al., 2007; Staginnus et al., 2007). Ces siARN interviendraient dans la régulation des EPRV (infectieux ou non) par un contrôle permanent de type 'gene silencing' au niveau transcriptionel (TGS) via la méthylation de l'ADN et/ou une modification des histones (Wassenegger, 2005), ou post-transcriptionnel (PTGS) par une dégradation site-spécifique des transcrits viraux issus des EPRV ou produits par des virus homologues (Almeida & Allshire, 2005; Baulcombe, 2004) (Figure 6a). Ce dernier mécanisme se traduirait par une résistance de la plante à l'infection par les virus génétiquement proches.

4.2.2 Activation des EPRV infectieux

Selon le modèle que nous venons de décrire, l'activation des EPRV associés à l'hétérochromatine ne pourrait résulter que du **dysfonctionnement du système de régulation** épigénétique de ces derniers (Figure 6b). Le **croisement génétique interspécifique** et les **stress abiotiques** sont les principaux facteurs activateurs identifiés et communs aux différents modèles d'EPRV infectieux (Dallot et al., 2001; Lockhart et al., 2000; Ndowora et al., 1999; Richert-Pöggeler et al., 2003) (voir partie 5.3.2 pour le BSV). Ces stress sont justement décrits comme induisant un relâchement du contrôle épigénétique, associé à la ré-activation des rétroéléments (Capy *et al.*, 2000; Grandbastien *et al.*, 2005; Sabot & Schulman, 2006; Slotkin & Martienssen, 2007).

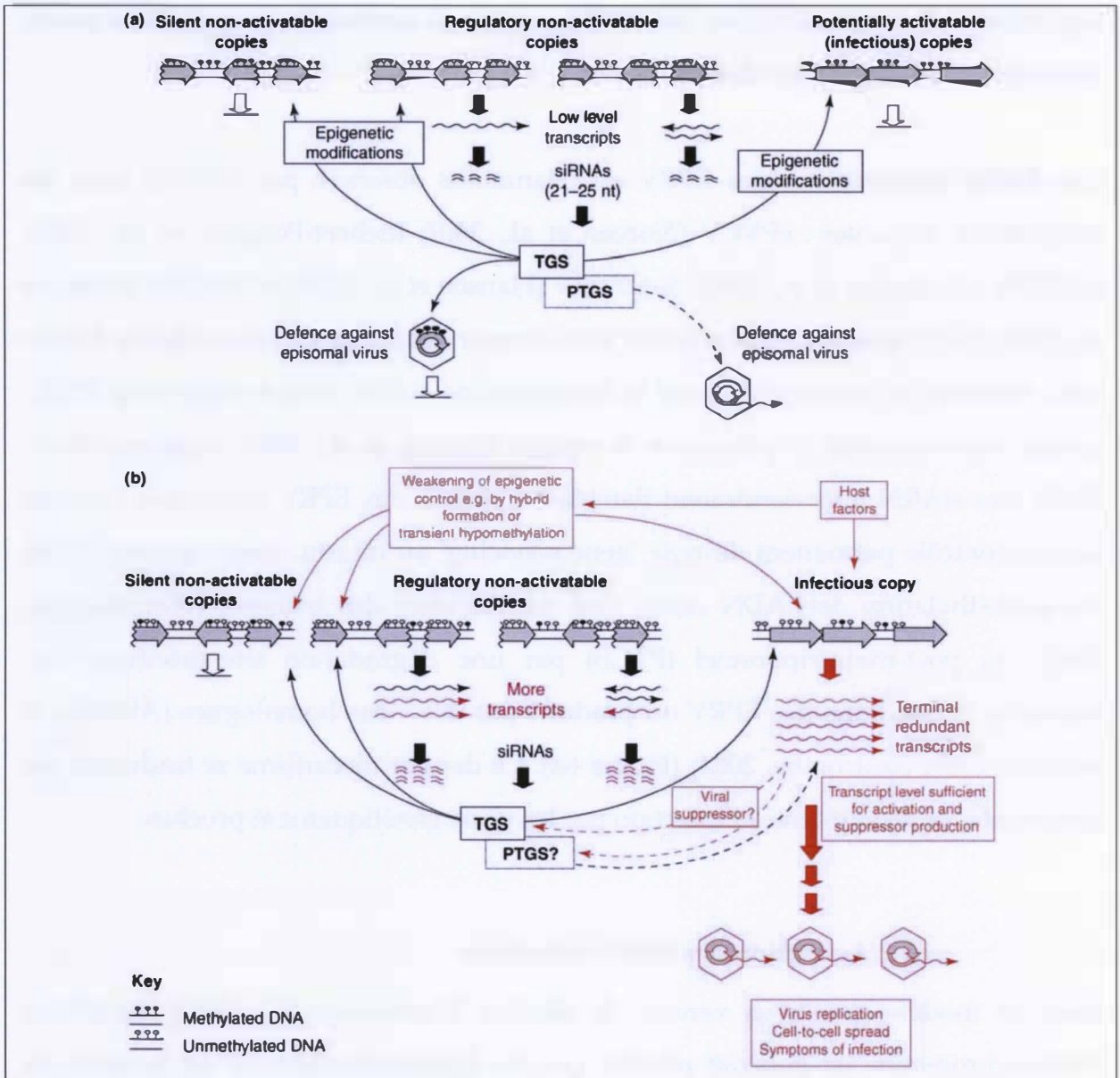

Figure 6 : Contrôle épigénétique des EPRV

(a) Modèle de regulation épigénétique des EPRV. Le génome des plantes contient un grand nombre de copies dites silencieuses car méthylées et associées à l'hétérochromatine (gauche), ainsi qu'un faible nombre de copies EPRV pouvant être transcrites à un faible niveau par l'ARN polymérase (milieu). La majorité des EPRV ont des ORF défectives, mais certains peuvent également être infectieux (droite). Les transcrits des EPRV sont utilisés comme matrice pour la production de siARN qui induisent ou maintiennent les modifications épigénétiques (méthylation de l'ADN et formation de l'hétérochromatine) des autres EPRV homologues par TGS (transcriptional gene silencing). Ce mécanisme permet également de réprimer l'activation des éventuels EPRV infectieux, ou des virus exogènes suffisamment proches. Des mécanismes de PTGS (posttranscriptional gene silencing – indiqués par les flèches en pointillé) pourraient également être impliqués.
(b) Modèle d'activation des EPRV. Le contrôle épigénétique qui permet en temps normal de réprimer l'activation des EPRV infectieux, peut être relâché lors d'une hypométhylation transitoire provoquée par une hybridation interspécifique, ou par d'autres facteurs internes. La transcription directe des EPRV disposés en tandem permet alors la production d'ARN génomiques viraux, conduisant à l'infection systémique par le virus. Adapté de Staginnus & Richert-Pöggeler (2006).

Ce modèle d'activation ne permet cependant pas d'expliquer l'hypothèse d'une activation des EPRV par recombinaison homologue au sein des EPRV, notamment pour les EPRV BSV pour lesquels ce dernier mécanisme est fortement suspecté (Ndowora et al., 1999).

4.2.3 Une résistance induite par les EPRV ?

Pour expliquer la présence et le maintien d'EPRV chez tant d'espèces végétales malgré leur potentiel pathogène, il a été fréquemment proposé que ces derniers soient associés à un **avantage sélectif** pour les plantes (Geering *et al.*, 2005a; Hohn *et al.*, 2008; Hull *et al.*, 2000; Staginnus & Richert-Pöggeler, 2006). Dans le modèle de contrôle génétique décrit dans la partie précédente, un niveau minimal de transcription des EPRV permet la production de siARN. Ces derniers maintiennent une transcription minimale des EPRV infectieux nécessaire au maintien de la régulation épigénétique (Figure 6a). En cas d'infection virale par des **virus épisomaux** homologues aux EPRV, la production de siARN doit permettre de lutter contre la multiplication virale par la dégradation des ARN viraux, dès lors que la séquence des virus est suffisamment proche de celle des EPRV. Cette **résistance de type ARNi** (ARN interférent) contre les infections des BSV épisomaux pourrait ainsi constituer cet avantage sélectif (Mette et al., 2002; Noreen et al., 2007; Staginnus et al., 2007).

Ce silencing basé sur les EPRV est fortement suspecté chez le pétunia, et pourrait également s'appliquer aux bananiers. Ce mécanisme expliquerait l'absence de multiplication virale observée chez les génotypes parentaux de tabac, de pétunia et de bananier porteurs d'EPRV infectieux (Hull et al., 2000; Lheureux, 2002; Lockhart et al., 2000; Mette et al., 2002). De même, chez les *Ns*EPRV, *Nto*EPRV, *Sotu*EPRV et *Lyc*EPRV, les virus épisomaux à l'origine des EPRV ne sont pas connus à l'heure actuelle. Outre un biais d'échantillonnage, ces populations virales auraient pu décliner voire s'éteindre complètement, du fait de mécanismes de résistance efficaces induits par les EPRV (Mette et al., 2002).

Ce type de résistance expliquerait le maintien des EPRV dans le génome des plantes, mais pas nécessairement le maintien de génomes viraux complets avec des ORF fonctionnelles. La résistance de type ARNi n'explique donc pas uniquement le maintien par sélection naturelle des EPRV infectieux (Hohn et al., 2008). Les mutations ponctuelles et les réarrangements dans la séquence des EPRV, observés chez tous les EPRV connus, peuvent en effet s'accumuler sans être contre-sélectionnés, tant qu'ils ne concernent pas les zones de production des siARN. Ces mêmes mutations pouvant en outre faire perdre leur potentiel pathogène aux EPRV infectieux, ces derniers deviendraient alors inoffensifs pour les plantes.

5 BSV-bananier : un modèle d'étude des EPRV infectieux

5.1 Les hôtes bananiers

Le premier partenaire du pathosystème étudié dans cette thèse sont les bananiers ; le seul hôte du BSV et le seul genre végétal connu pour posséder des EPRV BSV. Dans cette première partie, nous présenterons succintement les caractéristiques biologiques des bananiers, leur diversité spécifique et leur histoire évolutive, ainsi que les grandes étapes de leur domestication par l'homme. Ces informations seront en effet utiles lors de l'étude du polymorphisme d'intégration des EPRV en lien avec l'histoire évolutive des hôtes.

5.1.1 Description botanique

Dans l'ordre des *Zingiberales*, la famille des *Musaceae* ne comporte que deux genres. Les enset, **genre *Ensete***, pouvant atteindre 5 à 7 m de hauteur, ressemblent fortement aux bananiers (Figure 7). Ils possèdent 2n = 18 chromosomes. Six espèces ont été décrites (Simmonds, 1962) et sont cultivées en Ethiopie principalement (*E. ventricosum*), et constituent une partie importante de l'alimentation de 10 millions de personnes (Jones, 2000).

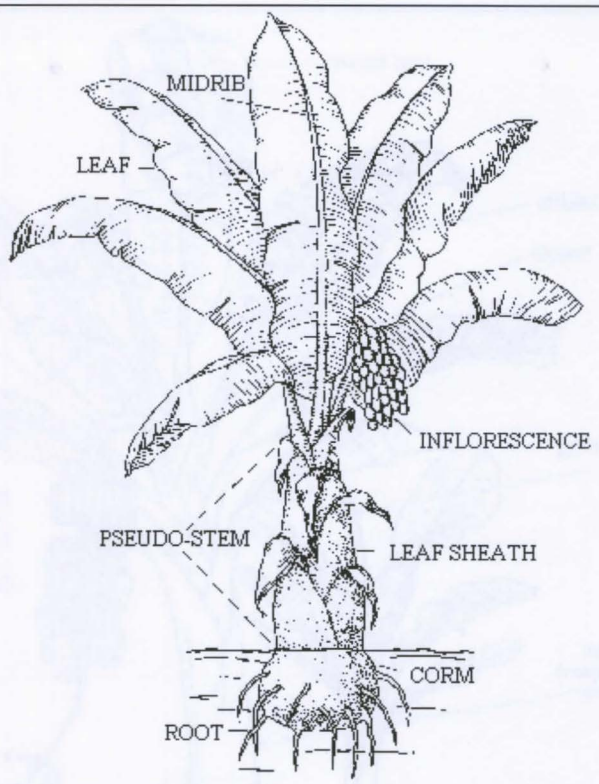

Figure 7 : Genre *Enset*.

Représentation de l'espèce Ensete ventricosum, adapté de Jones, (2000).

Les bananiers : **genre *Musa***, renferment les plus grandes plantes herbacées : si les bananiers cultivés font généralement 1,5 à 3 m de hauteur, certaines espèces sauvages atteignent 15 m. Les bananiers se composent d'une tige souterraine (appelée rhizome ou corme) d'où partent les racines, et d'un pseudotronc formé des gaines des feuilles serrées les unes aux autres (Figure 8). Le méristème apical, d'où partent les feuilles, est situé au centre du pseudotronc, à peu près au niveau du sol.

Les bananiers se **reproduisent de manière végétative** par rejetonnage. C'est au niveau de la tige souterraine que des bourgeons se développent, s'enracinent et forment des rejets. Les bananiers sauvages sont **séminifères** (*i.e.* fertiles) : la reproduction sexuée a lieu après émission de l'inflorescence. Chez certaines espèces, les fleurs sont hermaphrodites, mais en général, les fleurs mâles situées dans la partie apicale de l'inflorescence sont séparées des fleurs femelles situées dans la partie basale.

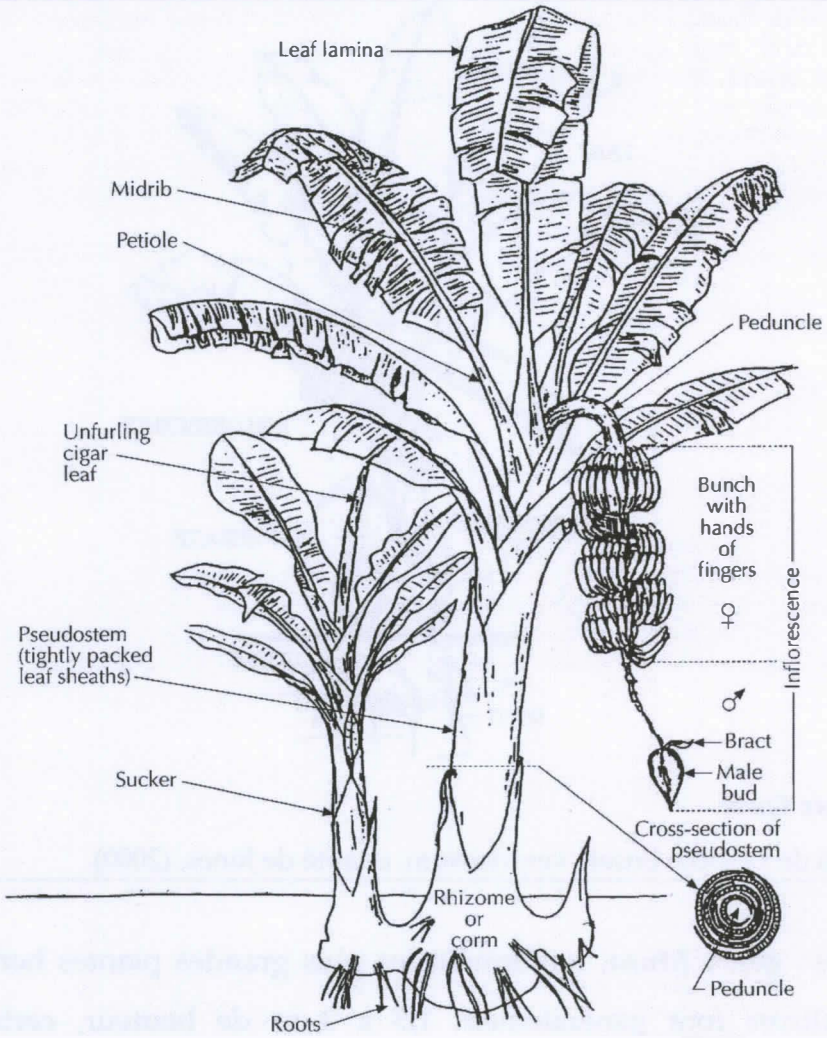

Figure 8 : Genre *Musa*

Représentation schématique d'un bananier en fleur et de son rejet, adapté de Jones, (2000).

5.1.2 Phylogénie et évolution du genre *Musa*

La distribution des bananiers sauvages se situe dans la zone intertropicale d'**Asie du Sud, du Sud-Est, et du Pacifique** (Simmonds, 1962), où ils occupent les vallées ou clairières humides mais bien drainées des forêts de faible et moyenne altitude.

Quatre sections ont été décrites sur des bases morphologiques et cytogénétiques : *Australimusa* et *Callimusa* à 2n = 20 chromosomes, et *Rhodochlamys* et *Eumusa* à 2n = 22 chromosomes (Cheesman, 1947). Plusieurs études phylogénétiques basées sur des marqueurs AFLP ou RAPD ont ensuite tenté de valider ces sections (Gawel et al.,

1992; Nwakanma et al., 2003; Ude et al., 2002; Wong et al., 2002). Il semblerait que les groupes 2n = 20 et 2n = 22 forment deux clades distincts, et que les sections à l'intérieur de ces deux groupes soient paraphylétiques. Ces travaux n'ont cependant pas réussi à clarifier de manière précise les relations phylogénétiques entre espèces de bananiers (Heslop-Harrison & Schwarzacher, 2007), notamment à cause d'un trop faible échantillonage, ainsi qu'à cause des marqueurs moléculaires utilisés.

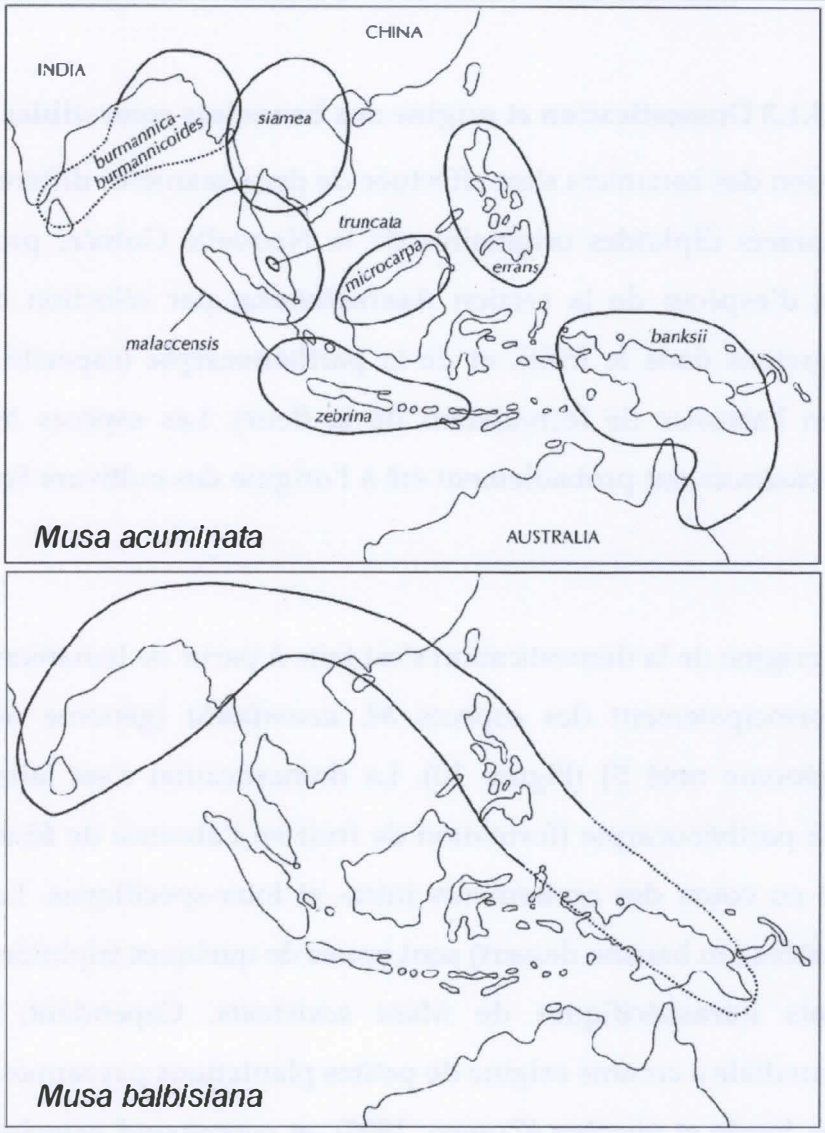

Figure 9 : Distribution naturelle des bananiers sauvages.

Aires approximatives de distribution des espèces *Musa acuminata* (en haut) et *Musa balbisiana* (en bas). Les sous-espèces de *M. acuminata* sont indiquées sur le schéma. L'espèce *M. balbisiana* est vraisemblablement originaire de l'Inde et du sud de la Chine, puis aurait envahi le sud-est de l'indonésie. D'après Simmonds, (1962) et Horry & Jay, (1997).

La distribution des deux espèces principales de la section *Eumusa* se situe en Asie du Sud-Est. *M. acuminata* est composée de neuf sous-espèces associées à des zones géographiques différentes (Figure 9A). Le centre d'origine de l'espèce *M. balbisiana* s'étend de l'Inde au sud de la Chine (Uma et al., 2005). Utilisée pour ses fibres, cette espèce aurait été introduite par l'homme lors de ses déplacements au nord de l'Indonésie et jusqu'en Papouasie Nouvelle Guinée (Argent, 1976). *M. balbisiana* serait redevenue sauvage dans ces pays (Figure 9B).

5.1.3 Domestication et origine des bananiers comestibles

La domestication des bananiers s'est effectuée de deux manières différentes. Les Fehi sont des bananiers diploïdes originaires de la Nouvelle Guinée, provenant de la domestication d'espèces de la section ***Australimusa***, par sélection de la **stérilité** (absence de graines dans le fruit), et de la **parthénocarpie** (capacité du fruit à se développer en l'absence de fécondation de la fleur). Les espèces *M. maclayi*, *M. peekelii* et *M. lolodensis* ont probablement été à l'origine des cultivars Fehi (Ude et al., 2002).

La deuxième origine de la domestication s'est faite à partir de bananiers de la section ***Eumusa***, et principalement des espèces ***M. acuminata*** (génome noté A) et ***M. balbisiana*** (génome noté B) (Figure 10). La domestication s'est faite d'abord par sélection de la parthénocarpie (formation de fruit en l'absence de fécondation) puis de la stérilité au cours des croisements intra- et inter-spécifiques. Les variétés de bananes exportées (ou banane dessert) sont issues de quelques triploïdes (AAA) issus de croisements intraspécifiques de *Musa acuminata*. Cependant, 87 % de la production mondiale a comme origine de petites plantations paysannes destinées à la consommation locale et vivrière (Gowen, 1995), et correspond principalement à des hybrides interspécifiques A x B, principalement triploïdes AAB et ABB. Des variétés diploïdes AA et AB existent et sont encore cultivées en Asie du Sud-Est (Simmonds & Shepherd, 1955).

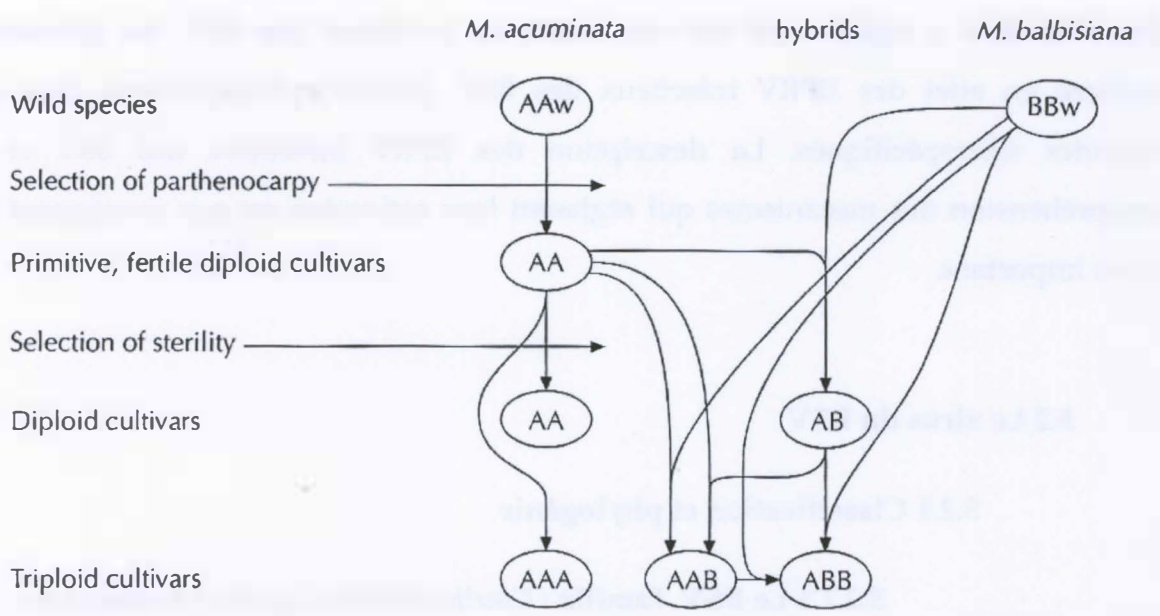

Figure 10 : Schéma de domestication des *Eumusa*

Evolution des principaux groupes de cultivars comestibles de bananiers par domestication des *Eumusa*. D'après Simmonds, (1962) et Carreel, (1994).

La découverte de nombreux phytolythes (cristaux typiques produits par les plantes) appartenant à des bananiers de la section *Eumusa* ont été découverts dans des sites archéologiques de Papouasie Nouvelle Guinée (Horrocks *et al.*, 2008; Lentfer & Green, 2004). Le nombre important de phytolythes retrouvés dans les sédiments analysés datant de 7000 ans, confirme que les bananiers ont été cultivés intensément à cette époque (Denham et al., 2003). Ces résultats permettent de retracer l'origine de la domestication des bananiers pendant le début de l'Holocène en Papouasie Nouvelle Guinée qui est l'un des 9 centres d'origine de l'agriculture (Neumann, 2003).

Aujourd'hui, l'amélioration variétale des bananiers vise à répondre à des contraintes agronomiques liées à l'émergence de parasites fongiques détruisant les cultures (cercosporioses), et à la nécessité de réduire les intrants phytosanitaires et les apports en eau dans les cultures de bananier. L'introgression de génome de *Musa balbisiana* dans des fonds génétiques *M. acuminata* par des croisements génétiques interspécifiques permet d'apporter les caractères recherchés (tolérance aux stress

hydrique, résistance aux maladies fongiques). Cependant, cette stratégie d'amélioration a rapidement été confrontée au problème des BSV. Le génome B contient en effet des EPRV infectieux des BSV, activés spécifiquement chez ces hybrides interspécifiques. La description des EPRV infectieux des BSV et la compréhension des mécanismes qui régissent leur activation est par conséquent un enjeu important.

5.2 Le virus du BSV

5.2.1 Classification et phylogénie

5.2.1.1 Le BSV, famille : *Caulimoviridae*, genre : *Badnavirus*

Les *Caulimoviridae* ont comme caractéristique principale l'utilisation d'une **réverse transcriptase** pour leur réplication. Ils sont également appelés Pararétrovirus de plantes, en rapport aux pararétrovirus d'animaux (où *Hepadnaviridae*), qui sont aussi des virus ADN double brin utilisant une réverse transcriptase. Cette dénomination fut utilisée pour distinguer les pararétrovirus des rétrovirus utilisant également une réverse transcriptase, mais dont le virion contient de l'ARN, et dont le transcrit ADN s'intègre dans le génome de l'hôte lors du cycle viral (cf. partie 2.1.3).

Les Pararétrovirus sont cependant un groupe polyphylétique et l'utilisation du terme tend à disparaitre. Les *Hepadnaviridae* sont en effet un groupe proche des *Retroviridae*, et les *Caulimoviridae* appartiennent à un groupe phylogénétiquement proche des *Metaviridae* qui sont des éléments transposables viraux, comme les célèbres retroéléments de type Ty3/Gypsy (Hansen & Heslop-Harrison, 2004; Malik & Eickbush, 2001) (Figure 11).

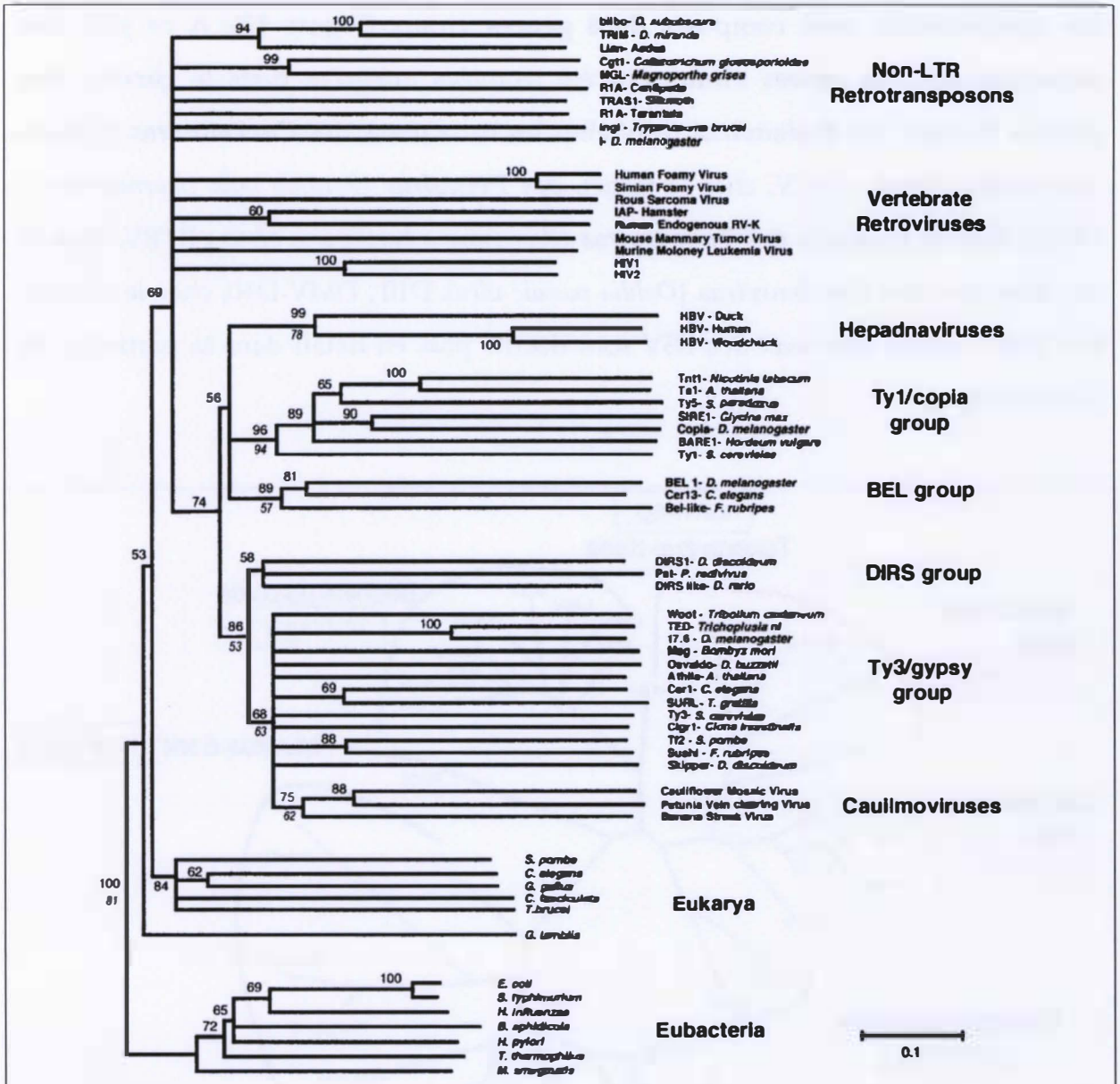

Figure 11 : Phylogénie des rétroéléments basée sur les domaines Ribonuclease HI (RNaseH)

Les eubactéries, eucaryotes et virus forment trois groupes bien distincts. Cette phylogénie révèle (1) la proximité phylogénétique entre les *Caulimoviridae* et les *Metaviridae* (groupe Ty3/gypsy) et (2) que les pararétrovirus (*Caulimoviridae* + *Hepadnaviridae*) sont un groupe polyphylétique. L'utilisation des domaines reverse transcriptase (240 aa) conduit à une topologie similaire, et confirme que les rétroviridae sont un groupe monophylétique et que ce groupe est proche du clade comprenant les *Hepadnavirida*, les *Metaviridae* et les *Caulimoviridae*. Dendrogramme reconstruit par Neighbor-Joining à partir de différents domaines RNaseH (140 aa). La robustesse des nœuds est testée par 1000 itérations de bootstrap indiquées à gauche des noeuds (multifurcation si < 50 %). Les domaines RNH des eubactéries sont utilisés comme groupe externe. Adapté de Malik et Eickbush, (2008).

Les *Caulimoviridae* sont composés de 6 **genres** viraux (Figure 12). A ce jour, des séquences de cinq genres viraux ont été trouvées intégrées dans le génome des plantes. Il s'agit des ***Badnavirus*** (BSV chez les bananiers), des ***Cavemovirus*** (*Tobacco vein clearing virus* ; TVCV, chez le tabac), des ***Petuvirus*** (*Petunia vein clearing virus* ; PVCV, chez le pétunia), des ***Tungrovirus*** (*Rice tungro bacilliform virus* ; RTBV, chez le riz) ainsi que des ***Caulimovirus*** (*Dahlia mosaic virus* D10 ; DMV-D10, chez le dahlia). Les EPRV autres que ceux des BSV sont décrits plus en détail dans la partie 4.1 de l'introduction.

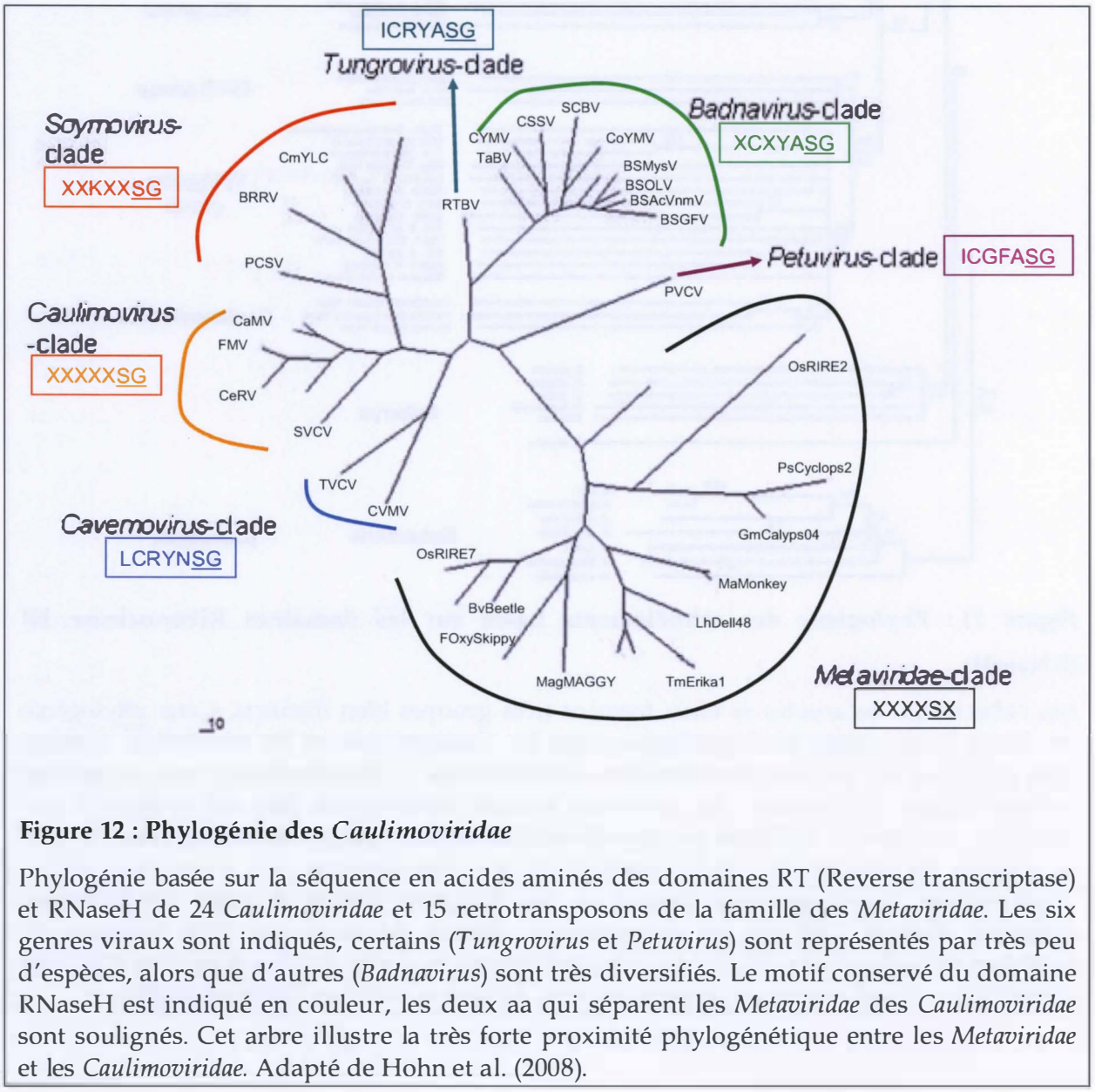

Figure 12 : Phylogénie des *Caulimoviridae*

Phylogénie basée sur la séquence en acides aminés des domaines RT (Reverse transcriptase) et RNaseH de 24 *Caulimoviridae* et 15 retrotransposons de la famille des *Metaviridae*. Les six genres viraux sont indiqués, certains (*Tungrovirus* et *Petuvirus*) sont représentés par très peu d'espèces, alors que d'autres (*Badnavirus*) sont très diversifiés. Le motif conservé du domaine RNaseH est indiqué en couleur, les deux aa qui séparent les *Metaviridae* des *Caulimoviridae* sont soulignés. Cet arbre illustre la très forte proximité phylogénétique entre les *Metaviridae* et les *Caulimoviridae*. Adapté de Hohn et al. (2008).

Les badnavirus infectent de nombreuses plantes tropicales, dont certaines cultures économiquement importantes comme la canne à sucre (*Saccharum officinarum*), la banane (*Musa* sp.), le cacao (*Theobroma cacao*), le taro (*Colocasia esculenta*) et les agrumes (*Citrus* sp.). La plupart d'entre eux sont transmis de manière **non circulante** par des cochenilles (Fauquet et al., 2005). Ce genre viral possède des particules virales bacilliformes non-enveloppées, d'environ 30 x 130 nm. L'organisation du génome viral est commune à tous les badnavirus, et comprend 3 ORF. Quelques badnavirus dérogent à cette règle et possèdent des ORF supplémentaires dont la fonction reste inconnue : comme le *Cacao swollen shoot virus* (CSSV) avec 5 ORF (Hagen et al., 1993) et le *Citrus mosaic virus* (CMBV) avec 6 ORF (Huang & Hartung, 2001).

5.2.1.2 Les BSVs, un complexe d'espèces virales

Le BSV a été décrit pour la première fois en Côte d'Ivoire en 1958 (Lockhart & Jones, 2000b) où il provoquait la 'mosaique en tiret' du bananier. Les premières études sur le BSV ont montré qu'il existe une très **forte hétérogénéité** entre les différents isolats, du point de vue sérologique, génétique et même biologique puisqu'ils induisent des symptômes différents sur la même plante (Lockhart & Olszewski, 1993).

Selon l'ICTV (International Committee on Taxonomy of Viruses), au moins un des critères suivant doit être observé afin de distinguer deux espèces du genre *badnavirus* : une différence dans la gamme d'hôte, une diversité nucléotidique dans la région RT/RNase H de l'ORF III d'au moins 20 %, une différence dans la séquence des produits des gènes, et une différence dans la spécificité des vecteurs (Fauquet et al., 2005). Des études de diversité génétique ont donc été réalisées à partir de cette région du génome viral, lors d'épidémies dans les cultures en Ouganda (Harper *et al.*, 2002; 2004; Harper *et al.*, 2005), en Australie (Geering et al., 2000) ou encore à l'île Maurice (Jaufeerally-Fakim et al., 2006). Ces travaux confirment que les différentes séquences de BSV que l'on croyait appartenir à des isolats génétiquement et sérologiquement diversifiés, correspondent en réalité à des **espèces virales distinctes**.

Cette très forte diversité génétique observée chez les BSV en particulier, et les badnavirus en général (Bousalem et al., 2008; Kenyon et al., 2008; Yang et al., 2003), pourrait provenir en partie, de l'**utilisation de la réverse transcriptase** pour la réplication du génome viral. En effet, cette ADN polymérase ARN-dépendante ne possède pas d'activité 'proof-reading', et fait **plus d'erreurs** que les ADN polymérase ADN-dépendant (Duffy et al., 2008; Flint et al., 2003).

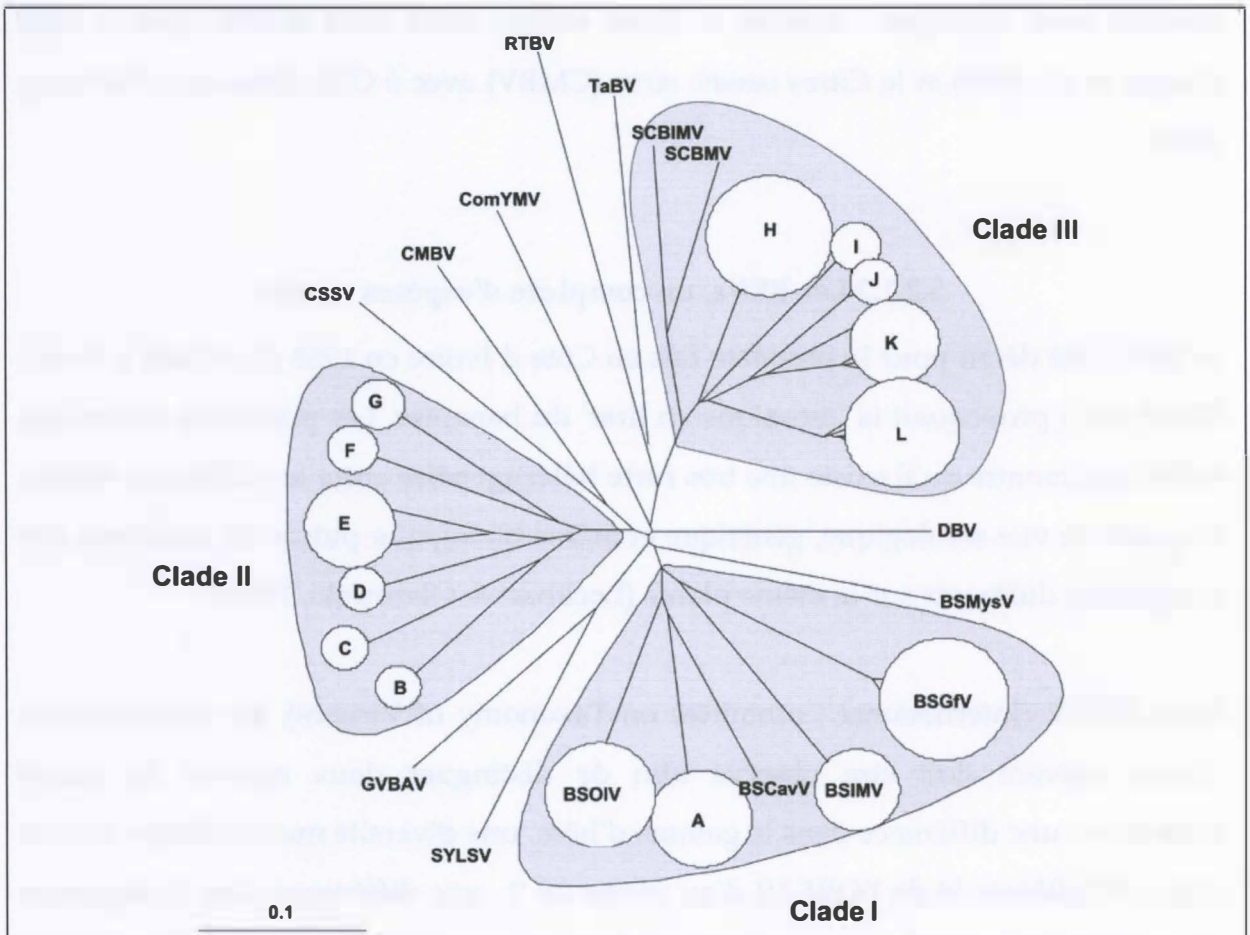

Figure 13 : Phylogénie des BSV

Dendrogramme non enraciné construit à partir de la région RT/RNAseH de l'ORFIII. La distinction de 3 'clades' (aire grisée) est basée sur une identité nucléotidique d'au moins 72 %. Les cercles blancs représentent les espèces décrites en Ouganda, soit trois espèces antérieurement connues (BSOLV, BSImV et BSGFV) ainsi que 12 potentielles nouvelles epèces (notées A à L). D'après Harper et al. (2005).

D'après les phylogénies basées sur la région de l'ORFIII comportant le domaine RT, les séquences des BSV ont divergé en **trois lignées** distinctes, tôt dans leur évolution (Figure 13). Un premier groupe comprend les premières espèces des BSV décrites : *Banana streak OL virus* (**BSOLV**) (Harper & Hull, 1998), *Banana streak Mysore virus* (**BSMyV**) (Geering et al., 2005b), *Banana streak Cavendish virus* (**BSCavV**) (Harper et al., 2005), *Banana streak Acuminata Vietnam virus* (**BSAcVNV**) (Lheureux et al., 2007), *Banana streak GF virus* (**BSGFV**) (Gayral et al., 2008) et *Banana streak Imové virus* (**BSImV**) (Gayral et al., In prep). Les deux autres groupes correspondent à des séquences nouvellement décrites en Ouganda sur des cultures de bananiers (Harper et al., 2005). D'autres espèces de badnavirus infectant diverses plantes hôtes auraient des relations phylogénétiques très étroites avec les BSV. Il s'agit par exemple du *Kalanchoe top-spotting virus* (KTSV) dans le groupe 1 (Geering et al., 2005a), du *Cacao swollen shoot virus* (CSSV) et du *Commelina yellow mottle virus* (ComYMV) dans le groupe 2 (Bousalem et al., 2008; Geering et al., 2005a), et du *Sugarcane bacilliform Mor virus* (SCBMV) dans le groupe 3 (Harper et al., 2005).

5.2.2 Particule virale et structure du génome

Comme tous les badnavirus, les BSV sont des virus non-enveloppés possédant des particules **bacilliformes** mesurant 120 à 150 x 30 nm (Figure 14).

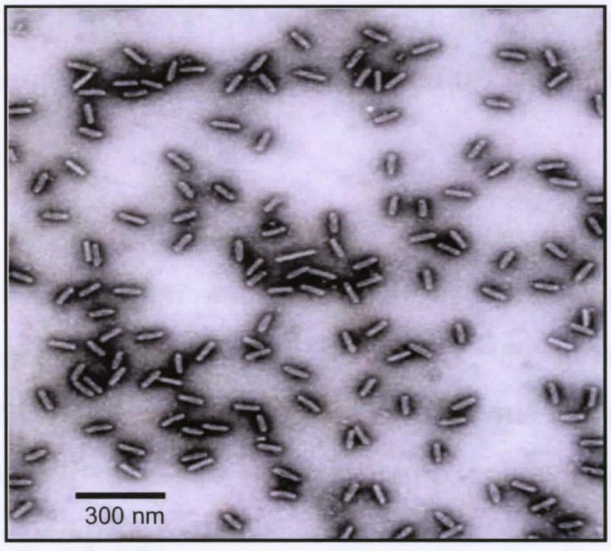

Figure 14 : Particule virale des BSV

Particules de BSV capturées sérologiquement avec l'anticorps polyclonal antiBSV IgG PMX2RC et observées en microscopie électronique. x 200 000.

Les BSV possèdent un génome ADN double brin (Lockhart, 1990) de 7 à 8 Kpb, comportant une **discontinuité** sur chacun des deux brins (voir les Figures 4 et 5) (Harper & Hull, 1998).

Le génome des BSV se compose de 3 cadres ouverts de lecture (ORF) (Figure 5). Les ORFI et ORFII codent deux petites protéines d'environ 22 et 14 kDa, respectivement (Harper & Hull, 1998), et dont la fonction reste mal connue. Cheng et al., (1996) ont montré que chez le ComYMV (*Commelina yellow mottle virus*), le produit de l'**ORFI** est associé à la fois aux virions et aux composants de la cellule hôte, ce qui suggèrerait qu'il participe à la formation des virions immatures. Le produit de l'**ORFII** est quant à lui associé aux virions matures, ce qui suggère un rôle dans l'assemblage du virus pour l'ORFII (Cheng et al., 1996). Chez le cacao swollen shoot virus (CSSV), le produit de l'ORFII possède une capacité de liaison aux acides nucléiques, notamment à l'ADN (Jacquot et al., 1996). Selon les auteurs, cette protéine pourrait intervenir au niveau de l'encapsidation de l'ADN génomique, voire dans le transport et la protection des ARN viraux des badnavirus. Enfin, le produit de l'**ORFIII** est une polyprotéine de 210 kDa. Par analogie avec le *Sugarcane bacilliform virus* (ScBV) (Bouhida et al., 1993) et le ComYMV (Medberry et al., 1990) qui sont des badnavirus proches des BSV, il est possible de déduire les fonctions de l'ORFIII par la présence de motifs conservés. Le clivage par protéolyse de cette polyprotéine restitue la protéine de capside virale (CP), l'aspartique protéase (AP), la réverse-transcriptase (RT), la RNAseH (RH) ainsi qu'une protéine putative de mouvement de cellules à cellules.

5.2.3 Biologie du BSV

Initialement considérés comme des souches d'une même espèce, les premiers isolats étudiés ont rapidement été élevés au rang d'espèces virales distinctes. Très récemment, deux travaux sur la diversité génétique des BSV libres et intégrés ont à nouveau bouleversés la vision que l'on avait sur ce virus (Geering et al., 2005a; Harper et al., 2005). En effet, plusieurs dizaines d'espèces potentielles de BSV existeraient, soit infectant à l'heure actuelle les bananiers, soit éteintes et représentées seulement par des EPRV. Très peu d'études se sont intéréssées à la biologie des BSV,

et aucune d'entre elles n'a comparé les caractéristiques des différentes espèces virales. C'est pourquoi, le terme générique BSV sera utilisé au singulier dans la partie suivante, et englobera les connaissances relatives aux différentes espèces connues de BSV.

5.2.3.1 Gamme d'hôte

Tout comme la plupart des badnavirus, le BSV possède une **gamme d'hôte** naturelle et expérimentale **très réduite**. Ces virus infectent naturellement les bananiers du genre *Musa* contenant du génome de *M. acuminata* (génotype AA, AAA, AB AAB, ABB, AAAB), et a été expérimentalement transmis par cochenille au genre *Ensete* (*E. ventricosum*) (Lockhart, 1995). Les essais d'inoculations du BSV via des cochenilles sur d'autres espèces du genre *Musa* comme *M. textilis* (Lockhart & Jones, 2000b) ou *M. balbisiana* n'ont pas abouti à une infection virale.

De manière intéressante, *M. balbisiana*, qui contient de nombreuses séquences intégrées du BSV dans son génome, semble être **résistant à la multiplication virale** quelle que soit l'origine **endogène** (EPRV) ou **exogène** (via cochenille) du virus (Iskra Caruana M.L. et al., 2003; Lheureux, 2002). Les hypothèses concernant la résistance des diploides *M. balbisiana* introduites dans la partie 4.2.3 seront discutées dans la partie 5.3.4.

Du fait de la gamme d'hôte très réduite, la transmission du BSV ne peut s'effectuer que de bananier à bananier ; les plantes adventices ne jouent donc aucun rôle direct dans l'épidémiologie du BSV, au contraire de ce qui est observé pour des virus infectant de nombreuses espèces végétales comme le *Cucumber mosaic virus* (CMV) (Lockhart & Jones, 2000a).

5.2.3.2 Symptômes et conséquences pour les bananiers

Les travaux sur le BSV ont débuté dans les années 1960, lorsque la maladie de la mosaique en tirets du bananier (BSD) a émergé au sein des cultures de bananiers en Afrique. D'une manière globale, l'infection par le BSV peut **réduire le rendement**

d'une parcelle de 5 à 15 %, bien que des taux moindres soient observés pour quelques cultivars plus tolérants (Dahal et al., 1998; Daniells et al., 2001).

Le BSV induit également des **symptômes chlorotiques foliaires** en tirets qui évoluent en **nécrose** (Figure 15A-B). Le diagnostic de la présence du BSV s'est pendant longtemps basé sur l'observation de ces symptômes, qui peuvent toutefois être absents même lorsque la plante est infectée par le virus. Des techniques plus précises permettant de détecter la présence du génome viral par PCR quantitative (Delanoy et al., 2003) ou de détecter les particules virales formées par immuno-capture-PCR (Harper et al., 1999a; Le Provost et al., 2006) sont depuis utilisées et couplées aux observations de particules bacilliformes par ISEM (immunosorbent electron microscopy).

Dans les cas les plus sévères, c'est-à-dire dans le cas d'une espèce agressive de BSV sur génotype sensible de bananier, l'infection peut aboutir à la **mort des plants** par nécrose du méristème apical conduisant à l'éclatement du pseudo-tronc (Figure 15C-D) (Lockhart & Jones, 2000b).

Plusieurs **stress abiotiques**, comme les fortes amplitudes de régime hydrique ou de température, semblent jouer un rôle sur l'expression des symptômes du BSV (Dahal et al., 1998; Daniells et al., 2001). Ces stress peuvent agir directement sur l'issue d'une infection virale déjà en place, mais peuvent également contribuer à l'activation d'EPRV infectieux (cf. partie 5.3.2.1).

Bien qu'aucune étude ne se soit penchée sur la prévalence et les conséquences des différentes espèces de BSV en populations naturelles, nous pouvons supposer, au vu des conséquences négatives qu'ils provoquent chez les bananiers cultivés, qu'ils puissent également affecter négativement les populations de bananiers sauvages en conditions naturelles.

Figure 15 : Symptômes de la maladie de mosaïque en tirets des bananiers

(A) Chloroses en tirets jaunes sur le limbe des feuilles. **(B)** Nécrose sur le limbe des feuilles. **(C)** Atteinte du méristème apical et mort de la feuille naissante (D) Eclatement du pseudo tronc à sa base, entrainant la mort du plant.

5.2.3.3 Cycle de réplication

Le cycle de réplication du BSV et des badnavirus n'est pas connu, et s'inspire de celui du *Cauliflower mosaic virus* (CaMV) (Figure 16), un virus appartenant au genre *Caulimovirus* (Jacquot et al., 1997) voisin du genre *Badnavirus*. Les principales étapes du cycle des BSV semblent être partagées avec celles du CaMV, à l'exception du nombre de transcrits (un seul pour le BSV) et de la forme minichromosome qui n'a pas été détectée pour le BSV (Lheureux, 2002).

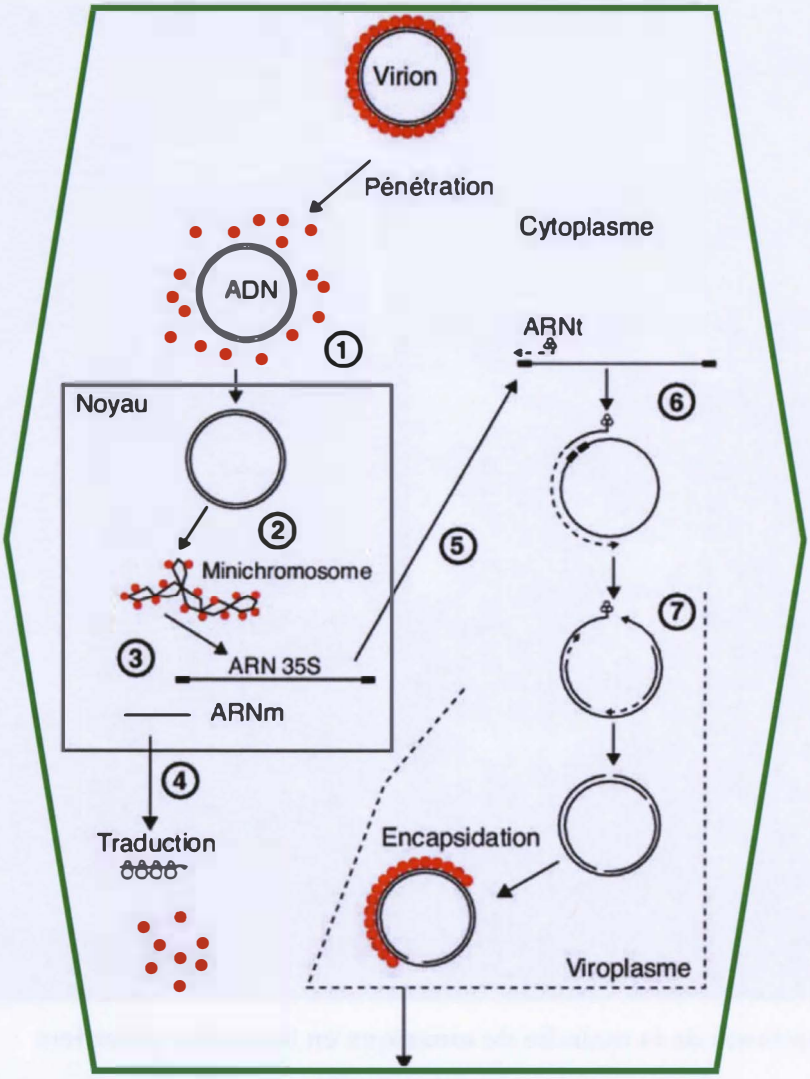

Figure 16 : Cycle de réplication des *Caulimovirus* (CaMV)

(1) Après la dissociation de la particule virale, l'ADN pénètre dans le noyau. **(2)** Les interruptions de séquence sont réparées à l'aide d'enzymes cellulaires et l'ADN viral est alors transformé en une molécule super-enroulée (minichromosome). **(3)** La transcription génère les ARN prégénomiques, dont le 35S de taille supérieure au génome. **(4)** Exportation des ARN dans le cytoplasme et traduction des protéines structurales et enzymatiques virales. **(5)** L'ARN prégénomique sert à la fois d'ARNm pour la synthèse des protéines virales et de matrice à la transcriptase inverse pour la synthèse du génome viral ADN. **(6)** L'amorçage de l'étape de transcription inverse s'effectue à l'aide de l'ARNt^{met} permettant la synthèse dans un premier temps d'un brin d'ADN (-). L'ARN matriciel est dégradé en petits ARN par la RNAse H, sauf au niveau de certaines régions riches en purines. **(7)** Ces petits ARN non dégradés sont utilisés comme amorces pour la synthèse des brins d'ADN complémentaires (+). Les protéines de capsides s'associent à l'ADN viral pour former de nouveaux virions.
D'après Hull, (2002), et Iskra-Caruana et al. (2003).

Le cycle débute par la **transcription** du génome ADN en ARN pré-génomique, qui joue un rôle de messager pour la production des protéines virales, mais qui sert

également de matrice pour la **transcriptase inverse** par laquelle l'ADN génomique viral est synthétisé.

L'intégration d'ADN viral dans le génome des bananiers résulte probablement de recombinaisons illégitimes (cf. partie 3.3) entre le minichromosome viral (s'il existe) formé dans le noyau, et le génome des bananiers.

5.2.3.4 Transmission du BSV

En contexte de culture, le BSV se propage facilement lors de la **multiplication végétative** naturelle des plantes, par l'utilisation des rejets, ou par la diffusion de vitroplants issus d'une plante contaminée. Ce mode de transmission permet la dissémination du BSV sur de longues distances, entre continents. Le virus n'est par ailleurs pas transmis mécaniquement sur bananier (Lockhart & Jones, 2000b).

Le deuxième mode de **transmission horizontale** du BSV se fait de manière non circulante par des **insectes vecteurs**. Quatre espèces de cochenilles (*Hemiptera*: *Coccoidea*: *Pseudococcidae*) infectant les bananiers sont capables de transmettre le virus : la cochenille des agrumes (*Planococcus citri*), la cochenille de la vigne (*P. ficus*), la cochenille rose de la canne à sucre (*Saccharicoccus sacœhari*) et la cochenille de l'ananas (*Dysmicoccus brevipes*) (Kubiriba et al., 2001; Meyer et al., 2008). Bien que la cochenille soit sédentaire, certains stades larvaires sont mobiles et peuvent ainsi assurer une transmission de plante à plante (Lockhart, 1995).

Des travaux antérieurs à la découverte des EPRV, ont montré qu'une **transmission verticale** du BSV peut également se faire par la **graine** (Daniells et al., 1995). Cette transmission est vraisemblablement liée à la présence d'**EPRV infectieux** dans le génome bananier producteur de graines. Les EPRV infectieux sont dès lors un mode efficace de **transmission verticale** du BSV (Hohn et al., 2008; Hull et al., 2000; Iskra-Caruana et al., 2003).

5.3 Les EPRV BSV

5.3.1 Découverte d'EPRV infectieux

L'amélioration génétique des bananiers a conduit à la création d'hybrides interspécifiques entre les espèces *M. acuminata* et *M. balbisiana.* Dans les années 1990, une **recrudescence de la maladie de la mosaïque en tirets des bananiers** (Banana streak disease - BSD) due à des infections par le BSV a été observée lors de la diffusion et de la mise au champ d'hybrides interspécifiques nouvellement créés (Folliot et al., 2005; Harper et al., 1999a; Lheureux, 2002). En 1996, la corrélation entre l'apparition de la maladie et la présence d'EPRV BSV mise en évidence par PCR et Southern blot dans le **génome B** (provenant de *M. balbisiana*), fut formellement énoncée (Lafleur et al., 1996). Le BSV est alors devenu la contrainte forte de la culture des bananiers. Plusieurs plantes peuvent en effet être infectées spontanément sur une même parcelle, rendant la gestion de ce pathogène difficile. Les EPRV infectieux sont donc responsables de l'expansion géographique du BSV par diffusion d'hybrides interspécifiques, si bien que le BSV est aujourd'hui le virus le plus répandu parmi les *Musaceae* cultivés (Daniells *et al.*, 2001; Geering *et al.*, 2000; Lockhart & Jones, 2000b).

Les premiers travaux sur les EPRV du BSV mettent en évidence l'intégration de l'espèce **BSOLV** dans le génome du bananier triploide Obino l'Ewai (génotype AAB) (Harper et al., 1999b). Les auteurs observent par FISH (Fluorescent *in situ* hybridization) un **faible nombre de sites** d'intégration correspondant à deux types de locus d'environ 50 et 150 Kb, respectivement. Chaque locus se compose de séquences virales répétées entre des séquences de plantes. Les auteurs ont pu séquencer au niveau d'un seul des locus, une portion seulement de la séquence virale intégrée. Cette partie d'EPRV s'avère être un réarrangement complexe de fragments du génome du BSOLV en sens opposés (Ndowora et al., 1999). Ce type d'intégrant est fortement suspecté d'être infectieux et de restituer des génomes fonctionnels du BSOLV observés chez ces hybrides en dehors de contamination extérieure possible par les insectes vecteurs.

Un deuxième cas de suspicion d'EPRV infectieux concerne le virus **BSMyV**. Curieusement, les cultivars 'Mysore' qui sont des hybrides interspécifiques de génotype AAB cultivés traditionnellement en Inde, possèdent de manière permanente un des symptômes foliaires de la maladie, initialement attribué à un désordre génétique (Wardlaw, 1972). Ce n'est que plus tard, suite à la caractérisation et au diagnostic des BSV, que ces symptômes ont été attribués à la présence de BSV (Lockhart, 1994) appartenant à une nouvelle espèce : BSMyV (Geering et al., 2005b).

C'est en étudiant le croisement génétique impliquant *M. balbisiana* cv. PKW ('Pisang Klutuk Wulung') diploïde de génotype BB, et *M. acuminata* cv. IDN110 4x, tétraploïde de génotype AAAA, que Lheureux et co-auteurs ont observé chez les hybrides AAB le réveil des espèces **BSOLV**, **BSGFV** et **BSImV** à partir d'EPRV présents dans le génome B du parent cv. PKW (Iskra Caruana M.L. et al., 2003; Lheureux, 2002; Lheureux et al., 2003). Une banque BAC a été réalisée à partir du cv. PKW (Safar et al., 2004) et analysée par hybridation moléculaire avec des sondes virales des espèces disponibles. Le séquençage des différents clones BAC portant les EPRV BSV est en cours, et les premières analyses de la structure de l'intégration de BSOLV chez ce cultivar montrent qu'il s'agit d'un EPRV différent de celui présent chez le cultivar Obino L'Ewai (Iskra-Caruana, non publié).

Malgré les avancées significatives sur l'étude des EPRV infectieux du BSV, le peu de locus d'intégrations caractérisés a déjà montré la diversité et complexité des intégrations. De plus, il n'existe à l'heure actuelle aucune démonstration formelle de la nature infectieuse des EPRV du BSV et les seules données publiées à ce jour se basent sur des corrélations entre la présence d'EPRV et l'apparition du BSV sans possibilité d'infection extérieure. Il est donc apparu crucial de faire avancer les recherches dans ce sens.

5.3.2 Activation des EPRV BSV

5.3.2.1 Facteurs déclencheurs

Un premier facteur déclencheur de l'activation des EPRV infectieux est l'**hybridation interspécifique** qui entraîne des bouleversements génomiques dus à la réunion de deux génomes différents et à la fréquente polyploïdisation associée. La majorité des hybrides bananiers d'origine naturelle ou nouvellement créés à des fins agronomiques, sont issus des croisements entre différents génotypes des espèces *M. acuminana* et *M. balbisiana* et possèdent une ploïdie variable. Plusieurs d'entre eux possèdent des EPRV infectieux apportés par le génome B, et sont capables de restituer spontanément des particules de BSV en dehors de toute transmission par cochenille. Il s'agit en particulier du cv. 'Obino l'Ewai' (AAB) (Ndowora et al., 1999), de l'hybride synthétique 'FHIA 21' (génotype AAAB) (Dallot et al., 2001), des cv. 'M'Bouroukou' et 'Kelong Mekintou' (AAB) (Folliot et al., données non publiées), de l'hybride CRBP 39 (AAAB) et des hybrides issus du croisement 'PKW x IDN110 4x' (AAB) (Lheureux, 2002).

Les **stress génomiques** constituent un facteur déclencheur nécessaire à l'activation des EPRV infectieux. Ce type de facteur est également en cause dans les deux autres pathosystèmes possédant des EPRV infectieux. Dans le cas du TVCV-tabac (cf. partie 4.1.2.1), l'activation est déclenchée chez les tabacs hybrides interspécifiques polyploïdes entre *N. clevelandii* et *N. glutinosa* ; dans le cas du PVCV-pétunia (partie 4.1.2.2), chez les pétunias polyploïdes hybrides entre *P. integrifolia* ssp. *inflata* et *P. axillaris* ssp. *axillari*. Dans ces deux exemples, le parent diploïde porteur des EPRV ne développe pas d'infection virale.

Chez les bananiers, la multiplication par **culture *in vitro*** (CIV) d'hybrides interspécifiques *M. acuminata* x *M. balbisiana* comme l'hybride AAAB cv. 'FHIA 21' (Dallot et al., 2001), ou AAB cv 'Dominico Harton' (Muller et al., données non publiées) provoque également l'activation des EPRV infectieux dès les premiers cycles de multiplication. Or, la culture cellulaire (Di Franco et al., 1992; McKenzie et al., 2002) provoque des stress qui induisent fortement la mobilisation des éléments transposables. Ces stress peuvent par exemple induire des modifications épigénétiques favorisant l'activation des rétroéléments (Capy *et al.*, 2000; Slotkin & Martienssen, 2007). Il est donc possible que la CIV des bananiers hybrides induise également un stress génomique qui favoriserait l'activation des EPRV infectieux.

Les **stress abiotiques** constituent le deuxième facteur déclencheur de l'activation des EPRV infectieux. Chez les bananiers hybrides interspécifiques, de fortes amplitudes thermiques et des écarts dans le régime hydrique ont été corrélées à des infections plus sévères par le BSV (cf. partie 5.2.3.2), mais le lien direct avec l'activation des EPRV infectieux n'a pas été formellement démontré.

Dans le cas des hybrides de pétunia ou de tabac, les stress abiotiques sont nécessaires au déclenchement d'une infection à partir des EPRV infectieux. Chez les pétunias, les tailles répétitives (Richert-Pöggeler et al., 2003), les stress hydriques (Lockhart & Lesemann, 1998), ainsi que les greffes et les chocs thermiques (Noreen et al., 2007; Zeidan et al., 2000) induisent une infection spontanée du PVCV.

Cependant, les mécanismes d'action des stress abiotiques dans l'activation des EPRV sont inconnus. A l'instar des stress génomiques, les stress abiotiques pourraient perturber la régulation des EPRV infectieux, comme cela a été montré pour l'activation des éléments transposables après des stress abiotiques (Capy *et al.*, 2000; Grandbastien, 1998; Slotkin & Martienssen, 2007). Mais ces stress abiotiques pourraient également affaiblir les plantes, conduisant à une défense moins efficace contre les infections virales débutantes provoquées par l'activation d'EPRV au niveau d'une cellule.

Enfin, la découverte de **facteurs génétiques** associés à l'activation ou au contrôle des EPRV infectieux laisse entrevoir une possibilité de comprendre les mécanismes génétiques de résistance au BSV chez *M. balbisiana* cv. PKW (BB), et éventuellement à plus long terme, d'utiliser cette résistance en amélioration variétale des bananiers. L'analyse de la ségrégation des marqueurs AFLP (Amplified Fragment Length Polymophism) dans les bananiers hybrides du croisement interspécifique *M. balbisiana* cv. PKW x *M. acuminata* cv. IDN110 4x (AAAA) a en effet montré l'existence d'un locus 'BSV expression locus' (BEL) associé à l'apparition du BSV (Lheureux et al., 2003). BEL est absent du génome du parent *M. acuminata* et présent à l'état hétérozygote chez le parent *M. balbisiana.* Dans la descendance, un des allèles BEL est retrouvé chez tous les descendants hybrides devenus infectés par le virus BSOLV, soit 50 % de la population. Les 50 % restants ne sont pas infectés par le BSV, et possèdent l'autre allèle BEL. Il serait intéressant de déterminer la nature exacte de

ce locus. En effet, BEL pourrait être une région régulatrice ou un gène régulant l'expression d'autres facteurs génétiques ou épigénétiques liés aux contrôles des EPRV.

5.3.2.2 Mécanismes d'activation des EPRV BSV

A ce jour, les connaissances sur les structures des EPRV BSV se limitent à un seul exemple, celui de l'intégration BSOLV dans le cv. 'Obino l'Ewai' (AAB) (Harper et al., 1999b; Ndowora et al., 1999). Malheureusement, les locus BSOLV EPRV sont nombreux dans ce génotype de bananier, et la séquence d'une partie d'un des locus seulement a pu être identifiée (environs 15 Kb). L'étude de cet EPRV s'est révélée néanmoins intéressante. Bien que présentant de nombreux réarrangements (fragments du génome viral tronqués, insertions, duplications), l'EPRV contient la totalité des informations génétiques nécessaires à la restitution d'un génome BSOLV complet. Au regard de cette structure complexe, les auteurs ont émis deux hypothèses d'activation de l'EPRV BSOLV (Figure 17). La première est basée sur **deux étapes de recombinaison homologue** permettant l'excision d'une molécule ADN circulaire ayant toutes les caractéristiques d'un génome viral fonctionnel. Dans la deuxième hypothèse, la première étape de recombinaison homologue est toujours nécessaire, mais elle est suivie d'une **transcription** de l'EPRV recombiné. La transcription aboutit à la formation d'un ARN pré-génomique infectieux qui participe au cycle de l'infection du BSV (décrit dans la partie 5.2.3.3).

Dans le cas de l'EPRV BSOLV, le locus est tellement réarrangé, qu'une à deux étapes de recombinaison homologue semblent nécessaires. Malheureusement, en l'absence de la séquence complète de l'EPRV étudiée et des séquences des autres locus dans le génome du cv. 'Obino l'Ewai', il est impossible de dire si cette hypothèse implique le scénario de recombinaison le plus parcimonieux. La proposition d'hypothèses et leur validation expérimentale sont donc très fortement dépendantes des connaissances sur le nombre de locus d'intégration dans le génome, ainsi que sur leur structure (degré de réarrangement de l'EPRV, présence ou non d'un génome viral complet, présence ou non de génomes répétés en tandem). Dans le cas du PVCV par exemple,

les intégrations dans le génome de *Petunia hybrida* sont en tandem et dans la même orientation (Richert-Pöggeler et al., 2003). Dans ce modèle, une transcription directe de l'EPRV semble être l'hypothèse la plus vraisemblable pour expliquer l'activation des ePVCV. Les auteurs ont par la suite également vérifié que les EPRV sont réellement transcrits (Noreen et al., 2007).

La **vérification expérimentale** des étapes d'activation des EPRV infectieux fait actuellement défaut dans le cas du BSV, et permettrait de valider les hypothèses sur les mécanismes d'activation.

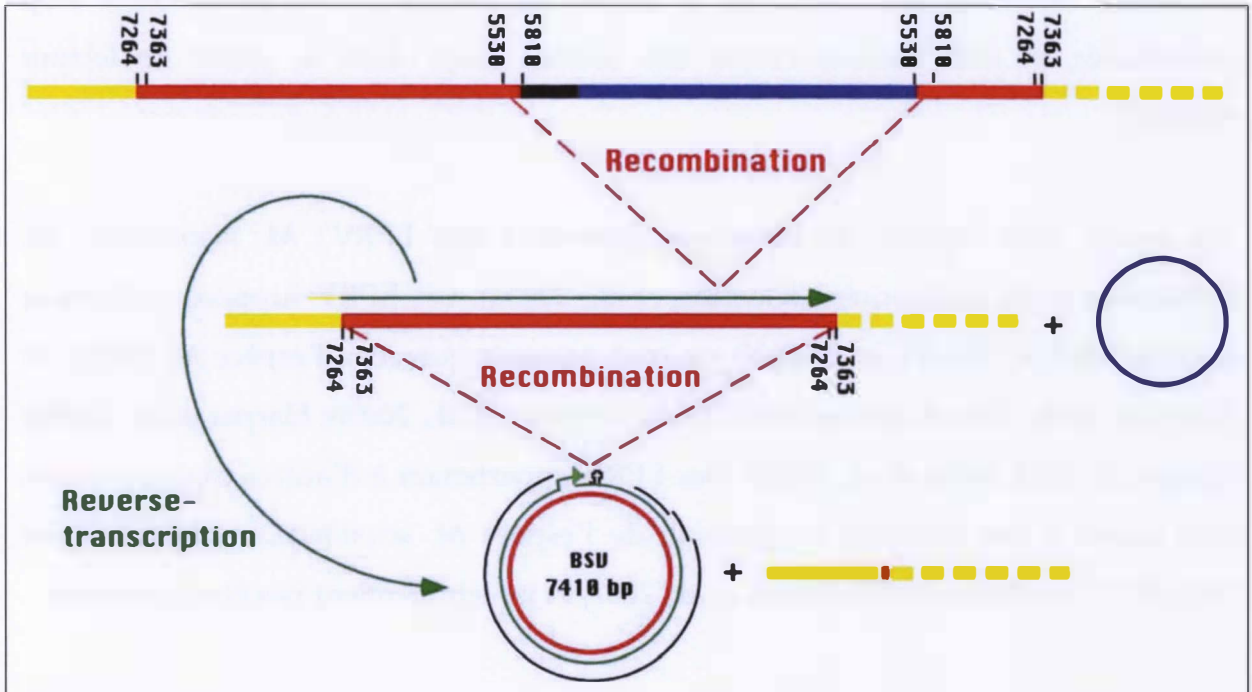

Figure 17 : Modèle d'activation de l'EPRV BSOLV chez le cv. AAB cv. 'Obino l'Ewai'

Les fragments de génome du BSOLV qui composent l'EPRV sont représentés en rouge (en bleu : sens opposé). Le génome du bananier est indiqué en jaune, en pointillé : séquence inconnue. En noir : séquences du génome *Musa* probablement retrouvées à l'intérieur de l'EPRV. La numérotation correspond à celle du virus BSOLV représenté en bas de la figure.

La première étape d'activation est une recombinaison homologue entre deux régions répétées de même orientation de 280 pb (représentées par les pointillés rouges). Cette recombinaison produit l'excision de la région centrale, et aboutit à un EPRV constitué d'un génome BSOLV linéaire, complet, et encadré par deux régions répétées de même orientation de 98 pb.

La deuxième étape conduit à la production d'un génome infectieux de BSOLV pouvant se faire par deux mécanismes. Le premier est une recombinaison homologue au niveau des séquences répétées (représentées par les pointillés rouges), qui aboutit à l'excision d'une molécule circulaire qui a la structure d'un génome BSOLV (en bas à gauche), et qui laisse un site vidé de d'EPRV (en bas à droite). Ce génome ADN est ensuite traduit en ARN pré-génomique qui initiera le cycle de réplication viral (cf. partie 5.2.3.3). Le deuxième mécanisme proposé est la transcription directe de l'EPRV recombiné (flèche verte), qui produira l'ARN pré-génomique. Adapté de Ndowora et al. (1999).

5.3.3 EPRV infectieux et non infectieux

Nos connaissances sur les séquences intégrées ont été grandement alimentées ces dernières années par les différents travaux menés sur les EPRV du BSV détectés par des approches d'amorces dégénérées ou par Southern blot. Ceux-ci ont révélé la présence de **très nombreux** 'variants' du BSV, intégrés dans **plusieurs espèces** de bananiers (Geering et al., 2005a). Bien que la séquence et la structure complète des locus d'intégration ne soient pas connues dans la majorité des cas, ces séquences obtenues sont très souvent **défectives**. Les ORF présentent fréquemment des mutations délétères (décalages de la phase de lecture créant des codons stop prématurés, et substitutions créant des codons stops dans la phase de lecture initiale).

Au moins trois espèces de bananiers possèdent des EPRV : ***M. acuminata, M. balbisiana*** et ***M. schizocarpa*** (Geering et al., 2005a). Les EPRV supposés infectieux (espèce BSOLV, BSGFV et BSImV) ne sont présents que chez l'espèce *M. balbisiana* (Geering et al., 2001; Geering et al., 2005a; Geering et al., 2005b; Harper et al., 1999b; Lheureux, 2002; Safar et al., 2004). Des EPRV appartenant à d'autres espèces virales sont quant à eux inféodés au génome de l'espèce *M. acuminata,* comme l'espèce BSAcVNV par exemple (Lheureux et al., 2007) et ne sembleraient pas être infectieux

Nous venons de voir que le génome des bananiers possède plusieurs EPRV infectieux, mais également de nombreuses séquences homologues d'EPRV défectives. Nous verrons dans la partie suivante que cette configuration n'est pas sans conséquences évolutives, tant pour les virus épisomaux infectant les bananiers, que pour les bananiers eux-mêmes.

5.3.4 Conséquences évolutives des EPRV BSV

5.3.4.1 Pour les virus

Chez le bananier, les EPRV du BSV sont en général **défectifs,** et ne sont par conséquent pas infectieux. Cette caractéristique est également observée chez les EPRV appartenant à d'autres modèles, comme les EPRV du RTBV chez le riz, ou les

EPRV chez les solanacées (Gregor et al., 2004; Jakowitsch et al., 1999; Kunii et al., 2004; Staginnus et al., 2007). Lorsqu'un bananier possédant des EPRV non infectieux est infecté par des BSV, il est possible que des échanges génétiques par recombinaison entre des virus épisomaux et leurs homologues intégrés se produisent. Sans être directement activés, les EPRV, même non infectieux, pourraient donc en théorie participer à la diversité génétique des virus épisomaux (Staginnus & Richert-Pöggeler, 2006).

Le cas des EPRV infectieux est naturellement différent. Bien qu'ils soient rares d'après la littérature, nous ne connaissons probablement pas l'étendue de ce phénomène tant il nous reste de génomes de plantes et d'espèces de *Caulimoviridae* à étudier. Les EPRV infectieux ont un réel impact dans l'**apparition et la transmission** horizontale des virus, et sont d'un point de vue évolutif, un acteur important dans les interactions hôtes-pathogènes des BSV, PVCV et TVCV avec leur hôtes respectifs.

En conditions expérimentales, les virus BSOLV et BSGFV provenant de l'activation des EPRV chez des hybrides interspécifiques de bananiers ont été **transmis horizontalement** à des plantes saines par plusieurs espèces de cochenilles (insectes vecteurs du BSV en condition naturelle) (Meyer et al., 2008). Il en est de même pour le BSMyV qui est fortement suspecté de provenir de l'activation d'EPRV infectieux chez des hybrides cv. 'Mysore' (génotype AAB) (Geering et al., 2005b). En conditions naturelles, les EPRV infectieux pourraient donc être **à l'origine d'épidémies** virales, et participer au **maintien de la diversité** du BSV au niveau intra et interspécifique. *In fine*, les EPRV infectieux pourraient limiter la probabilité d'extinction du virus à l'échelle de l'hôte ou des populations d'hôtes. Enfin, une activation simultanée d'EPRV peut conduire à la présence de plusieurs espèces de BSV dans un même bananier, pouvant ainsi favoriser des mécanismes de **recombinaisons entre virus** qui pourraient dès lors être à l'origine de l'émergence de nouveaux variants ou espèces virales.

Le revers de l'intégration, toujours du point de vue des virus, est que les mutations délétères qui sont normalement contre-sélectionnées au sein d'une population de virus épisomaux, ne le sont plus dans un génome viral intégré. Les EPRV subissent en effet l'évolution du génome de l'hôte, et ne peuvent qu'accumuler les mutations délétères les rendant **défectifs à long terme**. Les EPRV n'étant infectieux que lorsqu'ils sont récents (*i.e.* intégrés depuis peu), la 'survie' des virus intégrés au cours de l'évolution pourrait s'effectuer par des **évènements successifs d'intégration**, ou encore par des **échanges de matériel génétique** entre EPRV et BSV épisomaux lors d'une infection multiple, ou avec des EPRV non défectifs par recombinaison lors de la reproduction sexuée des bananiers *M. balbisiana*.

5.3.4.2 EPRV et évolution du génome des bananiers

La présence de nombreuses séquences EPRV du BSV dans le génome des bananiers participe dans une relative mesure à la **taille et à la complexité du génome** hôte. Il n'est cependant pas exclu qu'à plus long terme, les EPRV puissent s'amplifier et coloniser le génome des bananiers, de la même manière que ceux présents à l'heure actuelle en plusieurs milliers de copies dans les génome des *Solanaceae* (Gregor et al., 2004; Mette et al., 2002).

La présence de nombreuses copies d'EPRV peut également être à l'origine de modifications de la **méthylation** du génome, ainsi que de **réarrangements chromosomiques** majeurs au sein du génome (inversions, duplications, translocations…) (Hohn et al., 2008) de la même manière que peuvent le faire les éléments transposables dans les génomes eucaryotes (Bennetzen, 2000; Kidwell & Lisch, 2000).

5.3.4.3 Délétères et bénéfiques : le paradoxe des EPRV

Si l'on fait un parallèle avec les éléments transposables, il paraît assez probable que les nombreuses copies d'EPRV dans les génomes des plantes, qu'elles soient infectieuses ou non, puissent être associées à des **mutations d'insertion** délétères. C'est notamment le cas lorsque les EPRV s'intègrent dans ou près d'un gène et qu'ils

perturbent son expression ou modifient sa séquence codante. Cependant, chez le tabac et le pétunia, les locus d'intégration sont souvent les zones **les moins transcriptionnellement actives** du génome (ex. hétérochromatine, régions péricentromériques) (Hansen et al., 2005; Richert-Pöggeler et al., 2003; Staginnus et al., 2007). Ce biais d'intégration pourrait être la conséquence de **l'élimination par sélection naturelle** des intégrations les plus délétères voire léthales des zones transcriptionnellement actives du génome.

Bien que la majorité des EPRV BSV connus ne soient pas infectieux (Geering et al., 2005a), les quelques EPRV infectieux ont probablement eu un impact majeur sur l'évolution des bananiers. Suite aux premières intégrations infectieuses des BSV, les bananiers ont dû mettre en place des mécanismes moléculaires (nouveaux ou antérieurs aux intégrations), qui sont utilisés depuis dans la **résistance aux EPRV infectieux,** ces éléments 'cachés' dans le génome de leurs hôtes pouvant devenir délétères lors de leur activation. Les bananiers diploïdes *M. balbisiana* cv. PKW sont en effet résistants à l'activation des EPRV, puisqu'ils n'ont jamais été infectés par le BSV suite aux stress déclencheurs de l'activation des EPRV infectieux (CIV) alors qu'ils possèdent justement des EPRV infectieux (Lheureux et al., 2003). De plus, ces mêmes diploïdes *M. balbisiana* n'ont jamais été infectés par les BSV lors d'essais de transmission du virus par des cochenilles. Ces mécanismes de défense pourraient avoir pour origine une **régulation épigénétique** des EPRV (cf. partie 4.2), ou des **facteurs génétiques** identifiés (facteur BEL) (Lheureux et al., 2003), ou d'autres à identifier.

La défense contre les EPRV infectieux a cependant des limites, et elle semble être mise à mal par des stress génomiques notamment, comme l'hybridation interspécifique (cf. partie 5.3.2). L'activation des EPRV infectieux chez un bananier initialement sain induit une **infection par le BSV**. La virulence des espèces virales et la sensibilité et tolérance des génotypes de bananiers sont encore assez peu connues, mais dans l'ensemble, les infections par le BSV ont des effets néfastes pour le développement et la croissance de la plante (voir la partie 5.2.3.2). Etant donné que le génome B contient des EPRV infectieux de plusieurs espèces de BSV, il n'est pas rare

d'observer un réveil simultané de ces espèces (Folliot et al., 2005; Iskra Caruana M.L. et al., 2003; Lheureux, 2002).

Toutefois, chez les bananiers, seuls les hybrides interspécifiques sont sensibles à l'activation des EPRV infectieux. Les hybrides interspécifiques *M. acuminata* x *M. balbisiana* sont trouvés naturellement dans les zones de chevauchement des aires de répartition des deux espèces. Ces hybrides sont en général stériles et subissent donc une très forte dépression hybride. Le renforcement de cette dépression par l'activation des EPRV infectieux n'a par conséquent aucune conséquence négative sur les individus des espèces *M. acuminata* et *M. balbisiana*. Au contraire, cette dépression aurait même pu **faciliter la spéciation**. Du point de vue de la sélection naturelle, les EPRV infectieux ne sont pas directement délétères pour les individus *M. balbisiana*. Il serait cependant faux de dire qu'ils n'ont pas de coût. Le maintien des régulations épigénétiques et des résistances est en effet généralement associé à un fort **coût** qui se fera ressentir sur la fitness des bananiers notamment en termes d'allocation des ressources (Coustau et al., 2000).

6 Objectifs généraux de la thèse

Ce travail de thèse vise à comprendre le phénomène biologique original que sont les intégrations de séquences de *Caulimoviridae* chez les plantes, en prenant comme modèle biologique les EPRV du *Banana streak virus* présents dans le génome du bananier (*Musa* sp.).

Dans notre étude, nous nous sommes appliqués à étudier de manière conjointe les causes proximales (mécanistiques) et les causes ultimes (évolutives) associées aux EPRV, que nous présenterons successivement dans les deux chapitres suivants. Ces deux approches complémentaires deviennent indissociables dès lors que l'on souhaite comprendre le phénomène d'intégration virale dans son ensemble.

Dans le premier chapitre, nous avons tout d'abord décrit les intégrations infectieuses de l'espèce BSGFV dans le cultivar *M. balbisiana* cv. PKW, ainsi que leurs

particularités structurales et génétiques en lien avec leur potentiel d'activation (partie 1 - Article 1).
Nous avons ensuite étudié les mécanismes d'activation de ces EPRV infectieux et proposé un modèle théorique d'activation par recombinaison homologue, c'est-à-dire permettant la restitution d'un génome viral BSGFV fonctionnel à partir des séquences intégrées du BSGFV dans le cv. PKW. Ce modèle théorique a ensuite été validé expérimentalement (partie 2 - Article 2).

Le deuxième chapitre est consacré à l'évolution des séquences intégrées du BSV. Dans une première partie, nous avons réalisé une analyse phylogénétique des séquences du BSV en incluant pour la première fois les séquences épisomales et EPRV publiées. Cette étude à large échelle nous a permis de définir les relations phylogénétiques des BSV, de quantifier le phénomène d'intégration dans les bananiers, ainsi que de comparer les patrons d'évolution moléculaires entre les virus libres et intégrés (partie 1 - Article 3).
Enfin, nous avons suivi l'évolution de deux intégrations infectieuses de structure connue. La première est celle du BSGFV chez le cv. PKW étudiée dans le chapitre I ; la deuxième est celle du BSImV intégrée chez ce même cultivar. Dans cette analyse, nous avons étudié le polymorphisme d'insertion et l'évolution de la structure des EPRV depuis leur intégration. Afin de superposer l'évolution des plantes hôtes et celle des EPRV, nous avons reconstruit une phylogénie des espèces sauvages de bananiers qui possèdent cette intégration. Cette étude nous a permis de reconstruire la chronologie des intégrations et de leur évolution dans le génome des espèces du genre *Musa* (partie 2 - Article 4).

Les résultats obtenus au cours de cette thèse concernant les séquences intégrées du BSV chez les bananiers ont été valorisés par les publications suivantes :

Article 1

- P. Gayral, J.C. Noa-Carrazana, M. Lescot, F. Lheureux, B. E. L. Lockhart, T. Matsumoto, P. Piffanelli and M.L. Iskra-Caruana (2008). A single *Banana streak virus* integration event in the banana genome as the origin of infectious endogenous pararetrovirus. **Journal of Virology**, 82 (2), 6697-6710

Article 2

- P. Gayral, M. Royer and M.L. Iskra-Caruana. Evidence for activation of infectious endogenous pararetrovirus in banana (*Musa* sp.) by homologous recombination. Soumis à **Journal of General Virology**.

Article 3

- P. Gayral and M.L. Iskra-Caruana. Phylogeny of Banana streak virus reveals a recent burst of integrations in the genome of banana (*Musa* sp.). En reviewing à **Molecular Phylogenetics and Evolution**

Article 4

- P. Gayral, L. Blondin, O. Guidolin, F. Carreel, I. Hippolyte, X. Perrier and M.-L. Iskra-Caruana. Evolutionary history of infectious endogenous banana streak viruses and their host banana (*Musa* sp.). Soumis à **BMC Evolutionary Biology**.

Article 5

- P. Gayral, F. Lheureux, J.C. Noa-Carrazana, M. Lescot, P. Piffanelli, F. Carreel, C. Jenny and M.L. Iskra-Caruana. Exploring the banana streak viruses - *Musa* sp. pathosystem: how does it work? (2007), Proceedings of ISHS/ProMusa symposium White River, South Africa September 10-14. Recent advances in banana crop protection for sustainable production and improved livelihoods. **Acta Horticultura -** à paraitre. (Article en Annexe de la thèse).

Article 6

- M.L. Iskra-Caruana, P. Gayral, S. Galzi, N. Laboureau. How to control and prevent the spread of *banana streak virus* (BSV) when the origin could be viral sequences integrated in *Musa* genome? (2007), Proceedings of ISHS/ProMusa symposium White River, South Africa September 10-14. Recent advances in banana crop protection for sustainable production and improved livelihoods. **Acta Horticultura -** à paraitre. (Article en Annexe de la thèse).

Les données issues du séquençage et de l'annotation du clone BAC du bananier cv. PKW portant l'EPRV infectieux BSImV ont été analysées par Olivier Guidolin lors de son stage de master 2 que j'ai co-encadré. Cette étude a premièrement permis la construction de marqueurs moléculaires de l'intégration BSImV, qui ont été utilisés dans l'étude d'évolution des EPRV (Article 4).

La description de cet EPRV et son utilisation dans des expériences de bombardement ont ensuite été utilisées pour démontrer la nature infectieuse de l'EPRV. Ces résultats

en cours seront brièvement présentés en perspectives, mais ne seront pas présentés dans cette thèse. Ils font actuellement l'objet d'une publication en cours de rédaction (Article 7).

Article 7

- P. Gayral, O. Guidolin, J.C. Noa-Carrazana, P. Piffanelli, A. Geering, S. Sidibe-Bocs, F.C. Baurens, and M.L. Iskra-Caruana. Description of a new endogenous pararetrovirus species in the banana genome (*Musa sp.*) and evidence it is infectious. **In prep**.

Enfin, parallèlement à mon activité de recherche doctorale, j'ai participé aux travaux de thèse de Y. Abu-Ahmad en apportant une aide en phylogénie et évolution moléculaire dans le cadre d'une collaboration avec M. Royer (chercheur au CIRAD) et Ph. Rott (Directeur de thèse) et responsable de l'équipe « Génomique et analyse moléculaire de la pathogénie des bactéries phytopathogènes » de l'UMR BGPI. Cependant, pour conserver une unité, ces travaux ne seront pas présentés dans cette thèse.

L'étude porte sur l'analyse de la diversité génétique d'un virus hôte de la canne à sucre, le *Sugarcane yellow leafcurl virus* (SCYLV). Une nouvelle souche introduite sur l'île de La Réunion a été récemment décrite. L'étude de ses origines phylogénétiques a montré qu'il s'agissait d'une nouvelle espèce. L'analyse des mutations non-synonymes de son génome révèle la présence de plusieurs acides aminés évoluant sous sélection positive, probablement en lien avec un changement et une adaptation à son nouvel environnement. Les perspectives de cette étude s'ouvrent sur l'étude fonctionnelle des mutations identifiées : ont-elles un effet sur la gamme d'hôte, et jouent-elles sur la virulence du pathogène ? Ce travail a fait l'objet d'une publication co-écrite (Article 8).

Article 8

- M. Royer, Y. Abu Ahmad, P. Gayral, B. Moury, and P. Rott. Sequence analysis of Sugarcane yellow leaf virus isolates reveals evidence for a new virus species and evolutionary events. A soumettre.

Aims of the study

The aim of this work was the understanding of the original biological phenomenon of *Caulimoviridae* sequence integrations in the plants genome by taking as biological model the EPRV of *Banana streak virus* species Goldfinger integrated in the banana (*Musa* sp.) genome.

In our study, we were interested in bringing together the analysis of mechanism related to the activation of infectious EPRVs (proximal causes) and the understanding of evolution of EPRVs (ultimate causes), which will be presented in the next two chapters. These two complementary approaches become necessary since we wished to understand the phenomenon of viral integration in general.

In the first chapter, we described first of all the structure and genetic characteristics of the infectious EPRV of BSGFV species integrated in the cv. PKW, in correlation with their activation capacity (Part 1 – Article 1).
We then studied the mechanisms of activation of these infectious BSGFV EPRV. We proposed an hypothetical activation model based on homologous recombination, resulting in the release of a circular and complete functional viral genome from the BSGFV EPRVs of cv. PKW. This theoretical model was then experimentally validated (Part 2 – Article 2).

The second chapter treated the evolution of BSV EPRVs. In the first part, we realized a phylogenetic analysis of BSV sequences, including for the first time both episomal and integrated sequences available in public databases. This large scale study aimed to clarify the phylogenetic relationships among BSVs, to quantify the BSV integrations events into the banana genomes, and to compare the patterns of molecular evolution between episomal and integrated viral sequences (Part 1 – Article 3).
Finally, we followed the evolution of infectious EPRVs of two BSV species. BSGFV, described in the first chapter, and BSImV were both integrated into *Musa balbisiana* cv. PKW. The aim was to superimpose the evolution of the host plants genomes and

the evolution of EPRVs. To do so, we first inferred a phylogeny of the wild bananas harbouring these EPRVs. We then studied the polymorphisms of insertion and the evolution of the structure of EPRVs in relation to the host phylogeny. We could finally reconstruct the timing of integration events and the evolution stages of EPRVs into the genomes of species in *Musa* genus (Part 2 – Article 4).

Results obtained during this thesis concerning the BSV EPRV of banana were promoted by the following publications (Article 1-2-3-4-5).

Data resulting from the sequencing and annotation of the BAC clone of cv. PKW carrying the BSImV EPRV were analyzed by Olivier Guidolin (student in Master 2) that I co-supervised. Molecular markers specific of BSImV EPRV were first designed and used to study the EPRV evolution (Article 4). Then, to demonstrate this EPRV was infectious, it was cloned and used in biolistic delivery experiments (Article 7 in preparation).

Finally, outside of my PhD work, I collaborated with M. Royer and P. Rott on the PhD work of Y. Abu-Ahmad (Cirad) to define the phylogeny and molecular evolution using the genetic diversity of *Sugarcane yellow leafcurl virus* (SCYLV). This allowed us to describe a new viral species and positive selection was detected in this species recently introduced in La Réunion island (Article 8).

CHAPITRE I

Mécanismes et biologie des EPRV infectieux

Pour aborder les mécanismes responsables de l'intégration des séquences virales dans le génome du bananier, ainsi que de l'activation des EPRV infectieux du BSV, nous avons choisi de nous focaliser sur un modèle particulier d'intégration : les EPRV infectieux de l'espèce BSGFV, présents dans le bananier *Musa balbisiana* cv. PKW.

Ce premier chapitre est organisé en deux parties :

- la première concerne la description génomique et génétique des EPRV BSGFV dans le génome du cv. PKW, ainsi que la détermination de leur nature infectieuse (Article 1).

- dans la deuxième partie, nous proposons un modèle théorique d'activation des EPRV BSGFV par recombinaison homologue, suivi de la validation expérimentale de ce modèle (Article 2).

1 Structure et génétique de l'intégration infectieuse du *Banana streak GF virus* chez le bananier *Musa balbisiana* cv. PKW

1.1 Objectifs généraux

Cette étude a eu pour premier objectif la desctiption génomique des séquences intégrées du BSGFV dans le génome du bananier *M. balbisiana* cv. PKW. Le lien entre les capacités codantes des EPRV et les propriétés infectieuses de ces séquences a ensuite été établi par l'analyse de la structure et de la séquence des EPRV détectés.

Puis, nous avons déterminé la structure génétique des EPRV détectés, et vérifié si les EPRV BSGFV du cv. PKW sont infectieux, au travers de l'analyse de leur ségrégation dans le croisement génétique interspécifique utilisant PKW comme parent.

Enfin, nous proposons des hypothèses sur l'origine des intégrations du BSGFV dans ce bananier, en nous basant sur l'étude de l'environnement génomique proche des EPRV détectés.

Ces travaux sont présentés ci-après dans l'article publié dans la revue *Journal of Virology* :

P. Gayral, J.C. Noa-Carrazana, M. Lescot, F. Lheureux, B. E. L. Lockhart, T. Matsumoto, P. Piffanelli and M.L. Iskra-Caruana (2008). A single *Banana streak virus* integration event in the banana genome as the origin of infectious endogenous pararetrovirus. Journal of Virology, 82 (2), 6697-6710. Accepté le 7 avril 2008

1.2 Article 1 : "A single *Banana streak virus* integration event in the banana genome as the origin of infectious endogenous pararetrovirus"

JOURNAL OF VIROLOGY, July 2008, p. 6697–6710 Vol. 82, No. 13
0022-538X/08/$08.00+0 doi:10.1128/JVI.00212-08

A Single *Banana Streak Virus* Integration Event in the Banana Genome as the Origin of Infectious Endogenous Pararetrovirus▽

Philippe Gayral,[1] Juan-Carlos Noa-Carrazana,[2]† Magali Lescot,[2]‡ Fabrice Lheureux,[1] Benham E. L. Lockhart,[3] Takashi Matsumoto,[4] Pietro Piffanelli,[2]§ and Marie-Line Iskra-Caruana[1]*

CIRAD BIOS, UMR Biologie et Génétique des Interactions Plante-Parasite, TA 4-54/K Campus international de Baillarguet, F-34398 Montpellier Cedex 5, France[1]; CIRAD BIOS, UMR Développement et Amélioration des Plantes, Avenue Agropolis, TA40/03, FR-34398, Montpellier Cedex 5, France[2]; Department of Plant Pathology, University of Minnesota, St. Paul, Minnesota 55108[3]; and Plant Genome Research Unit, Division of Genome and Biodiversity Research National Institute of Agrobiological Sciences 2-1-2, Kannondai, Tsukuba, Ibaraki 305-8602, Japan[4]

Received 30 January 2008/Accepted 7 April 2008

Sequencing of plant nuclear genomes reveals the widespread presence of integrated viral sequences known as endogenous pararetroviruses (EPRVs). Banana is one of the three plant species known to harbor infectious EPRVs. *Musa balbisiana* carries integrated copies of *Banana streak virus* (BSV), which are infectious by releasing virions in interspecific hybrids. Here, we analyze the organization of the EPRV of BSV Goldfinger (BSGfV) present in the wild diploid *M. balbisiana* cv. Pisang Klutuk Wulung (PKW) revealed by the study of *Musa* bacterial artificial chromosome resources and interspecific genetic cross. cv. PKW contains two similar EPRVs of BSGfV. Genotyping of these integrants and studies of their segregation pattern show an allelic insertion. Despite the fact that integrated BSGfV has undergone extensive rearrangement, both EPRVs contain the full-length viral genome. The high degree of sequence conservation between the integrated and episomal form of the virus indicates a recent integration event; however, only one allele is infectious. Analysis of BSGfV EPRV segregation among an F1 population from an interspecific genetic cross revealed that these EPRV sequences correspond to two alleles originating from a single integration event. We describe here for the first time the full genomic and genetic organization of the two EPRVs of BSGfV present in cv. PKW in response to the challenge facing both scientists and breeders to identify and generate genetic resources free from BSV. We discuss the consequences of this unique host-pathogen interaction in terms of genetic and genomic plant defenses versus strategies of infectious BSGfV EPRVs.

Plant pararetroviruses are nonenveloped viruses with a noncovalently closed circular double-stranded DNA of 7 to 8 kbp (11). After infection, open circular viral DNA is released into the nucleus of the cell, where it is converted into supercoiled DNA and associates with histones to form a minichromosome. The viral DNA is then transcribed into mRNA, as well as pregenomic RNA, which is used for DNA replication in the cytoplasm via reverse transcription (23). In contrast to retroviruses, integration of the pararetroviral genome into the host genome is not required for viral replication. Nevertheless, pararetroviral integrations within the host genome exist and are assumed to have originated from illegitimate recombination during the minichromosome phase (53). Such integrants, called endogenous pararetroviruses (EPRVs), range from small, incomplete fragments to larger sequences, and become part of the plant genome via integration in a germinal cell subsequently becoming fixed in the plant population by the evolutionary forces of natural selection and/or genetic drift.

EPRVs are widespread within the plant kingdom. Thus far, the genomes of bitter orange (*Poncirus trifoliata*), potato (*Solanum tuberosum*), rice (*Oriza sativa*), tomato (*Lycopersicon* sp.), petunia (*Petunia* sp.), tobacco (*Nicotiana* sp.), and banana (*Musa* sp.) have been shown to harbor such integrants (24, 53). In 1999, Jakowitsch et al. (26) described tobacco EPRVs as a novel class of dispersed repetitive elements. EPRV can reach up to a 1,000 copies in tobacco (17, 37). The widespread distribution of EPRVs among plants, and their scattering within the host genome thus results in a discernible impact on host genome shape, plasticity, and evolution.

A surprising discovery was that some EPRVs could release virions. The data on the existence of these infectious EPRVs came from observations of spontaneous viral infection in petunia, tobacco, and banana by *Petunia vein clearing virus* (PVCV) (42), *Tobacco vein clearing virus* (TVCV) (34), and *Banana streak virus* (BSV) (7), respectively. The de novo apparition of these viruses followed stresses, wounding, or tissue culture processes in environments free of vector insects, suggesting that these viruses could only be derived from integrated

* Corresponding author. Mailing address: CIRAD BIOS, UMR BGPI, Campus International de Baillarguet, TA A-54/K, 34398 Montpellier Cedex 5, France. Phone: (33) 4 99 62 48 13. Fax: (33) 4 99 62 48 08. E-mail: marie-line.caruana@cirad.fr.

† Present address: Laboratorio Instituto de Biotecnologia y Ecologia Aplicada, Universidad Veracruzana, Av. Culturas Veracruzanas No 101, Col E. Zapata CP 91090, Xalapa, Ver., Mexico.

‡ Present address: Structural and Genomic Information Laboratory, CNRS UPR 2589, Institute of Structural Biology and Microbiology, Parc Scientifique de Luminy, 163 Avenue de Luminy, FR-13288 Marseille Cedex 9, France.

§ Present address: Rice Genomics Group, AgBiotech Research Centre, Parco Tecnologico Padano, Via Einstein, Località Cascina Codazza, 26900 Lodi, Italy.

▽ Published ahead of print on 16 April 2008.

forms. In 2003, Richert-Poggeler et al. (41) showed that PVCV EPRV (denoted ePVCV) is infectious by demonstrating release of a complete viral DNA genome that contributes to the viral infection. It is important to note that EPRVs, just like their exogenous counterparts, can lead to epidemics and are therefore of considerable economic importance.

BSV is a plant bacilliform pararetrovirus belonging to the family *Caulimoviridae* and the genus *Badnavirus* (22). BSV is one of five described viruses of banana (genus *Musa*) and plantain. This virus causes streak mosaic disease, which had until recently never been considered a serious threat (10). However, in the last 15 years, numerous spontaneous outbreaks of the disease have occurred in all banana-producing areas among promising banana breeding lines and micropropagated interspecific *Musa* hybrids, all originating from virus-free parents. The origin of these outbreaks was correlated with the presence of EPRVs in the genome of the cultivars. This phenomenon has contributed to the widespread distribution of BSV within banana-producing areas (33). Two types of BSV-related EPRVs have been described thus far in banana. The first type is defined by noninfectious sequences with nonfunctional viral open reading frames (ORFs) containing premature stop codons, frameshift mutations, and/or incomplete viral genomes. Such BSV EPRVs are present in the two most common *Musa* species from which most cultivated banana is derived: *Musa acuminata* (denoted A) and *Musa balbisiana* (denoted B) (16). BSV EPRVs of the other, so-called infectious type contain the complete functional viral genome.

The first tentative description of an infectious BSV EPRV concerns the 5′ part of the integrated species BSV Obino l'Ewai (BSOlV EPRV) present in the genome of the plantain cv. Obino l'Ewai (AAB) (38). This BSV EPRV has a complex structure consisting of noncontiguous back-to-back viral sequences, interrupted by *Musa* sequences. Although this BSV integrant is not fully described, it contains the entire BSOlV genome at least once. The authors of that study suspected BSOlV EPRV to be pathogenic and hypothesized a mechanism involving two homologous recombination events to release an infectious BSV genome.

Four natural widespread BSV species have thus far been identified as integrants: *Banana streak Obino l'Ewai virus* (BSOlV), *Banana streak Imové virus* (BSImV), *Banana streak Mysore virus*, and *Banana streak Goldfinger virus* (BSGfV) (16). In banana, abiotic stresses such as micropropagation by in vitro culture processes (7) and genetic hybridization (30) are known to contribute to triggering the production of episomal BSV from EPRVs. Studies on the apparition of BSV after interspecific genetic crosses revealed that at least two factors are involved in BSV expression. The first is the ploidy of the B genome in *Musa* genotypes. *M. balbisiana* diploid genotypes (BB) such as cv. Pisang Klutuk Wulung (PKW) and cv. Pisang Batu, which are used as female parents, harbor infectious BSV EPRVs in their genome but are nevertheless resistant to any multiplication of BSV, whether from EPRV activation or from exogenous BSV infection (25, 32). In contrast, genotypes with haploid B genomes harboring BSV EPRVs, such as the triploid hybrids (AAB) arising from interspecific genetic crosses, as well as other natural triploids (AAB cv. Kelong Mekintou and Black Penkelon) (12) or newly created tetraploids (AAAB FHIA 21) (7), can express BSV after stresses and are susceptible to BSV infection. The second factor is a genetic factor called BSV expressed locus (BEL) identified in the triploid (AAB) progeny of interspecific genetic crosses between virus-free diploid *M. balbisiana* (BB) cv. PKW and tetraploid *M. acuminata* (AAAA) cv. IDN 110 4x parents (30). In that study the authors characterized the segregation of BSOlV appearance among AAB F1 progeny expressing the disease as a Mendelian monogenic allelic system, strongly regulated by BEL and conferring the role of carrier on the *M. balbisiana* diploid parent, cv. PKW.

Comparisons with other well-described infectious EPRVs, e.g., PVCV in petunia and TVCV in tobacco, has unfortunately not been very informative up to now in suggesting ways to efficiently manage BSV expression. EPRVs differ considerably in copy number per genome and structure, as well as in their mechanisms of regulation by the host plant. For instance, EPRV expression is repressed by DNA methylation in petunia (39) and tomato (52), whereas this is not the case for BSV EPRVs (M. L. Iskra-Caruana, unpublished data).

Of the three latter pathosystems, BSV/*Musa* remains the most critical in terms of economic impact. Bananas are the developing world's fourth most important food crop, and three major issues concern BSV EPRVs. First, the main method of propagating banana plantlets is micropropagation by in vitro culture, which can trigger activation of BSV EPRVs. Second, in tropical zones global warming is responsible for strong variations of water regime and thermal amplitude, a well-established activator stress for BSV EPRVs (6). Finally, the numerous infectious BSV EPRVs of different BSV species are restricted to the B genome used in *Musa* breeding programs as a source of genes of agronomic interest. This consequently reduces considerably the possibility of using genetics to control banana sigatoka leaf spots, the main fungal constraint for the banana crop industry.

Until now, the description of a complete BSV EPRV and a more detailed analysis of the mechanisms of the activation of infectious EPRV have been lacking. To further describe the genetic mechanisms of the regulation of BSV EPRV, we report here the full molecular organization of the pathogenic BSV Goldfinger species (BSGfV) EPRVs in the genome of the wild diploid (BB) *M. balbisiana* cv. PKW and demonstrate that its integration is the result of a single event.

MATERIALS AND METHODS

Hybridization of BAC clones with BSGfV probe. Bacterial artificial chromosome (BAC) librarieswere obtained from the *M. balbisiana* wild diploid cv. PKW (43) and two *M. acuminata* banana plants: the wild diploid cv. Calcutta 4 (AA) (58) and the triploid "Cavendish" subgroup cv. Petite Naine (AAA). These BAC libraries were constructed by partial digestion of genomic DNA with HindIII restriction enzyme and cloning into the pIndigoBAC-5 HindIII cloning-ready and pCC1BAC HindIII cloning-ready vector. BAC DNA was isolated, digested to completion with NotI, and separated, alongside the lambda ladder PFG marker (New England Biolabs, Pickering, Ont.), by PFGE on a 1% (wt/vol) agarose gel in 0.5X TBE under the following conditions: 6 V/cm, switch time 5 to 15 s, and an angle of 120°C for 5 h at 14°C. Clones of the BAC libraries were spotted onto high-density Hybond N+ filters (AP Biotech, Little Chalfont, United Kingdom) by using a Flexys robot. The filters were hybridized with two BSGfV probes: full-length (pCR-TOPO [6,001 bp]) and fragment BSGfV (pCR-TOPO [1,262 bp]), covering the entire viral genome.

Fingerprint: digestion by restriction enzymes and use of Southern blotting. BAC DNA was digested with five different enzymes (HindIII, EcoRI, BamHI, PstI, and XhoI) to release the BAC fragments. The digested clones were separated on a 0.8% agarose gel in 1× Tris-acetate-EDTA at 60 V, run for 20 h. The

separated fragments were denatured and transferred to nitrocellulose membrane Hybond-N+ (Amersham Pharmacia Biotech) (45). Southern hybridization was realized in high-stringency conditions using both full-length or fragments of virus genome probes (45). Filters with the digested BAC clones were hybridized with the two BSGfV probes (pCR-TOPO [1,262 bp] and pCR-TOPO [6,001 bp]).

Sequencing of BAC clones. Selected BAC clones were sequenced by using the shotgun approach at the National Institute of Agrobiological Sciences. BAC shotgun sequencing was performed by using 2,000 shotgun (2-kb and 5- to 7-kb) clones of 10× coverage and a BigDye terminator kit (ABI) on ABI 3700 sequencers, assembled with Phred/phrap software (8, 9); contig gaps were filled by the primer-extension method when necessary. The GenBank accession numbers were AP009325 and AP009326 for MBP_71C19 and MBP_94116, respectively.

Sequence annotation. Each BAC sequence was processed through algorithms for predicting genes (FgenesH for monocot plants [44]; Softberry Software) and Genemark.hmm (35). The BLAST algorithm (1) was used for homology searches against nucleotide and protein databases. Information obtained by the different similarity searches and by the gene prediction programs was imported into the annotation platform Artemis (3) for further manual analysis. Dotter (51), REPuter (28), OligoRep (50; http://wwwmgs.bionet.nsc.ru/mgs/programs/oligorep/), and RepeatMasker (http://repeatmasker.org/) were used to search for repeated sequences. Gene structures and names were manually inspected and refined as necessary. Annotated gene models were scanned for *Musa* transposable element nucleotide sequences downloaded from GenBank. The BSV integrants sequences were manually annotated based on BSV sequences available in public databases.

Pairwise sequence comparison. Sequences were aligned with the CLUSTAL W algorithm (56) implemented in BioEdit (18) and corrected manually. Insertion and deletion events were removed prior to nucleotide identity calculation.

Interspecific genetic crosses of cv. PKW with cv. IDN 110 4x. The plant population used in the present study consisted of 165 F1 allotriploid hybrids (AAB) derived from interspecific genetic crosses between the virus-free wild diploid (BB) *M. balbisiana* female parent cv. PKW and the virus-free autotetraploid (AAAA) *M. acuminata* male parent cv. IDN 110 4x confirmed by immunosorbent electron microscopy and by immunocapture PCR (IC-PCR) (30). This genetic cross was fully described and characterized in Lheureux et al. (30). A total of 13% of the progeny was propagated in a vegetative manner to produce duplicates or triplicates of the original hybrid (235 hybrids). Leaf samples were stored at −80°C.

DNA extraction. Total DNA was extracted by the method described in Gawel and Jarret (14) from leaf tissue of AAB progeny stored at −80°C. The quality and amount of DNA was visually estimated after separation of 5 μl of DNA extraction in a 0.8% agarose gel, staining with ethidium bromide, and visualizing the sample with a UV transluminator.

PCRs. All PCRs were performed on 5 to 20 ng of template DNA using a common mix composed of 20 mM Tris-HCl (pH 8.4), 50 mM KCl, 0.1 mM each deoxynucleoside triphosphate, 1.5 mM $MgCl_2$, 400 nM of the forward and reverse primers, and 1 U of *Taq* DNA polymerase (Eurogentec, Seraing, Belgium) in a final volume of 25 μl. DNA was amplified after one cycle at 94°C for 4 min, 35 cycles of 94°C for 30 s, primer annealing at the temperature indicated for 30 s, 72°C for 1 min per kb, and a final extension at 72°C for 10 min. Amplicons were visualized after migration of 8 μl of PCR products on a 1.5% agarose gel in 0.5× TBE (45 mM Tris-borate, 1 mM EDTA [pH 8]). The gel was stained with ethidium bromide, and amplified bands were visualized under UV light.

EPRV genotyping. For the PCR-restriction fragment length polymorphism (RFLP) DifGf-TaaI method, the primers DifGfF (5′-TTGCAGGAGCAGGAA TTACA-3′) and DifGfR (5′-GGATGGAAGATGAGCTCTTTG-3′) (annealing temperature [T_a] = 60°C) amplify both ORF1 and ORF2 regions in BSGfV EPRVs (positions 702 to 1372 in BSGfV AY493509). PCR products (7 μl; 0.2 to 1.5 μg of DNA) were digested with 5 U of TaaI (Fermentas; restriction site 5′-AC,N′GT-3′) in 1× Tango buffer (Fermentas; 33 mM Tris-acetate [pH 7.9], 10 mM magnesium acetate, 66 mM potassium acetate, 0.1 mg of bovine serum albumin/ml) in a final volume of 10 μl. Incubations were performed at 65°C for 2 h. Digested DNA was loaded onto a 2.5% Nusieve (3:1) agarose (Lonza) gel stained with ethidium bromide, and the bands were visualized under UV light. In the multiplex PCR (VV3F/R-VV5F/R) method, the first set of primers (VV3F 5′-TTGCCAAGAATTCCTCCAAG-3′ and VV3R 5′-AAGTTCTTGTCGGCA AGGTG-3′, T_a = 60°C, positions 524 to 543 and 2888 to 2907 in BSGfV) hybridize with both alleles and yield an amplicon of 376 bp. The second set of primers (VV5F 5′-CCATGGAGGTTGACCTGTCT-3′ and VV5R 5′-ACCCC TCTGTCTTCCCAACT-3′, T_a = 60°C, positions 1896 to 1915 and 205 to 224 in BSGfV) hybridize with EPRV-9 only and yield a 628-bp amplification product. The multiplex PCR method generates a 1,012-bp product from the combination of primers VV5F and VV3R that hybridized elsewhere in both allelic EPRVs. In the PCR spe7/spe9bis method, we designed a set of PCR markers specific for *Musa* flanking regions of EPRV. Sequences of BAC MBP_71C19 and MBP_94116 were aligned by using CLUSTAL W (56). Insertion and deletion events were detected manually and then used to design PCR primers. A first set of primers located upstream of the viral integration site is specific to EPRV-7 (spe7F [5′-TGGCTACTCGTTTGCCTTTT-3′] and spe7R [5′-CCGTAGCTCT TGTGGCTAGG-3′], T_a = 59°C). A second set is specific to EPRV-9 and is located downstream of the EPRV (spe9bisF [5′-TGATAGAAATACTAAAGA TAGCTCATTACA-3′] and spe9bisR [5′-TTTTTGATTATTGCTTCTCTTTT T-3′], T_a = 50°C).

BSV genotyping: IC-multiplex-PCR-RFLP DifGf/Actin-TaaI. For BSV genotyping, the immunocapture step consisted of coating sterile polypropylene thin-walled 0.2-ml microfuge tubes (Axygen, Union City, CA) for 4 h at 37°C with 25 μl of immunoglobulin G purified from the polyclonal antiserum raised against BSV species and *Sugarcane bacilliform virus* species, diluted at 2 μg/ml in carbonate coating buffer (15 mM sodium carbonate, 34 mM sodium bicarbonate [pH 9.6]). The tubes were then washed three times with 100 μl of PBT washing buffer (136 mM NaCl, 1.4 mM KH_2PO_4, 2.6 mM KCl, 8 mM NA_2HPO_4, 0.05% Tween 20 [pH 7.4]). Plant extracts were prepared by grinding 0.5-g leaf samples in 5 ml of grinding buffer (2% polyvinylpyrrolidone 40, 0.2% sodium sulfite, and 0.2% bovine serum albumin prepared in PBT) using a manual bead grinder and plastic grinding bags (Bio-Rad Phytodiagnostics, Marnes-la-Coquette, France). Portions (1 ml) of plant extracts were transferred to microfuge tubes and clarified by centrifugation at room temperature for 5 min at 7,000 rpm. Then, 25 μl of the supernatant was loaded into coated tubes, followed by incubation for 1 h 30 min at room temperature. The tubes were washed five times with 100 μl of PBT, three times with 100 μl of sterile water, and then dried briefly. Multiplex PCR was carried out directly in tubes using DifGfF and DifGfR primers described above and Actine1F (5′-TCCTTTCGCTCTATGCCAGT-3′) and Actine1R (5′-GCC CATCGGGAAGTTCATAG-3′) primers that amplify the *Musa* actin housekeeping gene. PCRs were performed as described above at a T_a of 58°C for 25 cycles. A nested PCR using the internal primers VV1F (5′-ACAGCTCCAGG AGATTGGAA-3′) and GfM2 (5′- GAGCTCTTTGAGTCGTCATTG-3′) was then performed at a T_a of 63°C on 2 μl of diluted PCR products (1:100 or 1:1,000). TaaI digestion was subsequently carried out as described above on VV1F-GfM2 PCR products.

RESULTS

Detection of BSGfV EPRVs in the *Musa* nuclear genome. In order to exhaustively define the integration patterns of BSGfV in the banana nuclear genome, we used three BAC libraries derived from common cultivars representative of *Musa* cultivated species: the triploid *M. acuminata* Cavendish subgroup cv. Petite Naine (AAA), the wild diploid *M. acuminata* cv. Calcutta 4 (AA), and the wild diploid *M. balbisiana* cv. PKW (BB). We screened these three BAC libraries for integrated BSGfV by hybridizing with two viral probes covering the complete BSGfV genome. The *M. acuminata* libraries did not hybridize with either BSGfV probe. Of more than 36,864 BAC clones from the *M. balbisiana* library, 9 were found to contain BSGfV EPRVs (Table 1). Since the cv. PKW BAC library represents nine genome equivalents of *M. balbisiana* (43), the very small number of hits indicates low-copy integration.

We analyzed the nine BAC clones by using an RFLP and a fingerprint approach in order to establish whether they correspond to different integration events. Two BSGfV probes were used for hybridization. Among the five restriction enzymes tested, HindIII and PstI yielded informative patterns (Fig. 1) and led to the identification of two classes of BSGfV EPRVs in the genome of cv. PKW. We performed physical mapping and contig building of the nine positive BAC clones from the XhoI fingerprint by using FPC software (Fig. 2). All BAC clones cluster in a single contig with a high confidence level (score = 0.991), suggesting a single locus of BSGfV integration in the cv.

TABLE 1. Screen of *Musa* genomic BAC libraries by hybridization with BSGfV probes

BSGfV probe (size [bp])	*M. balbisiana* cv. PKW (BB)		*M. acuminata* (no. of hits)	
	No. of hits	BAC clones	cv. Calcutta 4 (AA)	cv. Petite Naine (AAA)
pPCR-TOPO (6,001)	9	30F18, 41K09, 48D15, 64H02, 71C19, 72M20, 73C24, 94116, 96J15	0	0
pPCR-TOPO (1,262)	9	30F18, 41K09, 48D15, 64H02, 71C19, 72M20, 73C24, 94116, 96J15	0	0

PKW genome. This result was subsequently confirmed by using a PstI fingerprint (data not shown). We removed incomplete EPRVs by discarding clones with BSGfV EPRV in the boundary regions. For this purpose, we sequenced the ends of each BAC clone (Table 2). Clones MBP_71C19 and MBP_94116 were selected for sequencing since they were representative of each of the two classes and contained the full BSGfV EPRV present within the BAC. The corresponding GenBank accession numbers are AP009325 and AP009326, respectively.

Structure of BSGfV EPRV-7 and EPRV-9. The sizes of the sequenced BAC clones MBP_71C19 and MBP_94116 are 133,041 and 119,244 bp, respectively. Each carries one copy of the full BSGfV integrant. Figure 3 shows the annotation of EPRV-7 and EPRV-9, named according to their BAC number. The integrants are much longer than a single BSGfV genome (7.26 kb): 13.28 kb for EPRV-7 and 15.58 kb for EPRV-9. The integrant is composed only of viral sequences, with no *Musa* genome sequences embedded within it. BSGfV EPRVs exhibit a complex rearrangement of viral sequences in the same and opposite orientation relative to the organization of the BSGfV genome. EPRV-7 is comprised of six juxtaposed viral fragments (I to VI), while EPRV-9 carries an additional segment. Both EPRVs are strikingly similar to each other since the first four fragments (I to IV) and the last fragment (VI) display the same structure and size. Although the EPRVs appear as a succession of several fragmented, inversed, and partially repeated BSGfV genomes, Fig. 4 shows that most of the viral regions in EPRV-7 and EPRV-9 (69 and 72%, respectively) are duplicated and therefore present in two or three copies within each EPRV. Most importantly, the entire BSGfV ge-

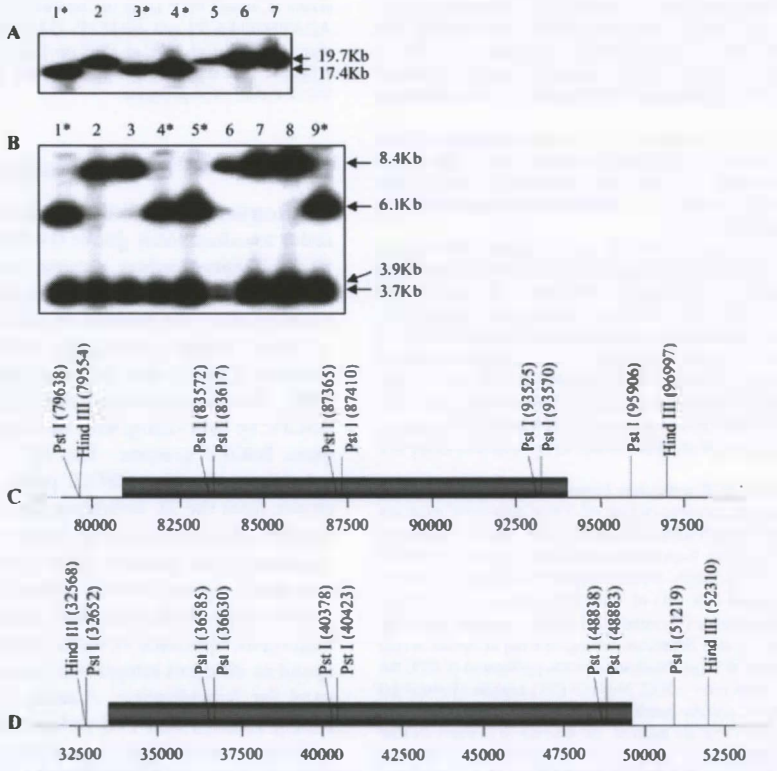

FIG. 1. Integration patterns of BSGfV EPRVs in cv. PKW. Fingerprint patterns obtained after digestion of BAC clones containing BSGfV EPRVs with HindIII (A) or PstI (B) and hybridization with two BSGfV probes covering the full-length viral genome are shown. Lanes 1 to 9 show the results obtained with BAC clones containing BSGfV inserts: lane 1, MBP 30_F18; lane 2, MBP 41_K09; lane 3, MBP 64_H02; lane 4, MBP 71_C19; lane 5, MBP 72_M20; lane 6, MBP 73_C24; lane 7, MBP 94_116; lane 8, MBP 48_D15; and lane 9, MBP 96_J15. Asterisks indicate BAC clones with the same restriction pattern. Deduced restriction maps of BSGfV EPRV in BAC clones MBP 71_C19 (C) and MBP 94_116 (D) are presented. Numbers refer to the position according to BAC annotation. Gray bars indicate the BSGfV EPRV.

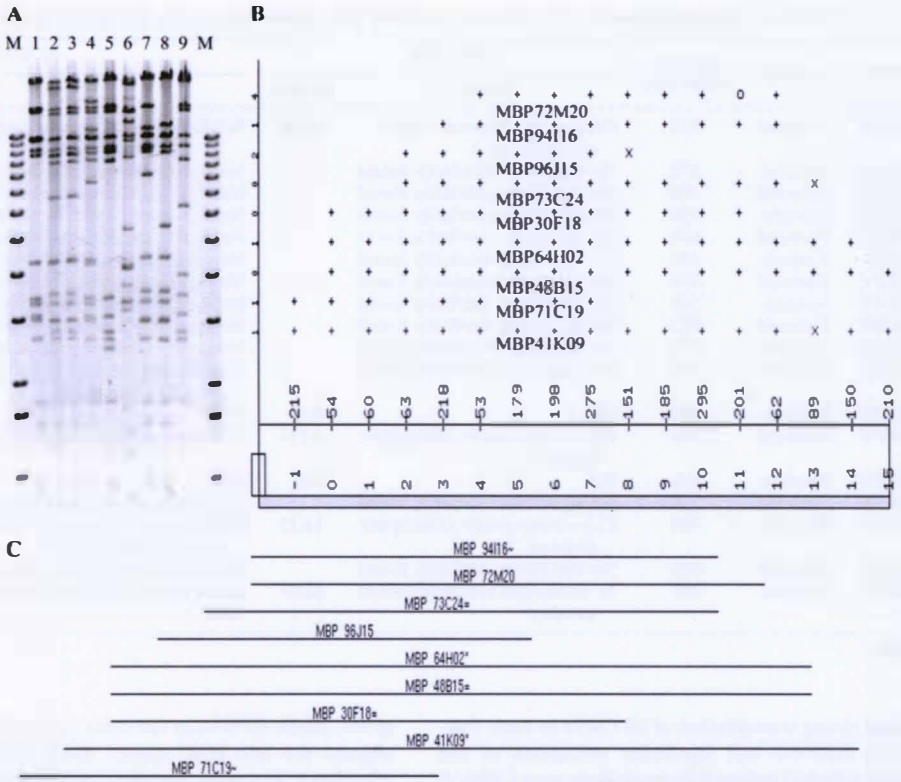

FIG. 2. Fingerprint contig building. (A) Restriction patterns of nine XhoI-digested BAC clones containing BSGfV integration analyzed with the software Image, version 3.10 (54). Lane M, 1-kb ladder (Invitrogen); lanes 1 to 9, BAC clones (the lane order is the same as in Fig. 1). (B) Consensus band map displayed in FPC version 4.7.9 (49) from the fingerprint analysis in panel A showing the ordering of clones and their fragments. At the top of the panel, the length of each clone is equal to the number of bands in the clone (total length of 17, ranging from −1 to 15). The points represent partially ordered groups: "+" indicates a match with the bottom band within the tolerance, "×" indicates a match within twice the tolerance, and "○" indicates no match. The middle portion of the panel indicates the consensus band numbers. The parameters used were as follows: tolerance, 7; cutoff, 10e-7. (C) Resulting contig of the nine BAC clones. The suffix symbols "*", "=", and "~" represent the status of each clone: the "*" indicates a parent clone, which shares the same common bands with exact child clones (=) or a percentage of the common bands with approximate child clones (~).

nome (the ORFs and intergenic region) is present at least once in each EPRV.

We then tested whether the high degree of similarity observed within and between EPRVs is confirmed by a similarity in nucleotide sequence. We performed a pairwise comparison of sequences within each EPRV. We chose a sequence of 1,847 bp present in two copies in EPRV-7 and in three copies in EPRV-9, thereby allowing intra- and inter-EPRV comparisons. This fragment is representative of the BSGfV genome since it contains part of the intergenic region, ORF1, ORF2, and the first 748 bp of ORF3 (positions 245 to 2090 of the BSGfV genome). The percentage of nucleotide sequence identity is very high within EPRV-7 (100%) and EPRV-9 (99.7 to 99.8%), as well as between EPRV-7 and EPRV-9 (99.8%) (Table 3). To support this result, we aligned the full sequence of EPRV-7 and EPRV-9 and compared the 13.1 kb of sequences they have in common. Only 28 substitutions differentiate EPRV-7 and EPRV-9, which thus share 99.8% nucleotide sequence identity.

Mutations accumulated in BSGfV EPRVs suggest that EPRV-7 is involved in release of functional BSGfV genomes. We analyzed the type of mutations that have accumulated in EPRV-7 and EPRV-9 (Table 4). Each EPRV fragment was compared to its corresponding homologous region in the BSGfV genome. Despite the very close similarity (99.3% identity on average) with the BSGfV sequence, the few differences in the EPRV sequences are relevant in terms of functional BSGfV virus released after EPRV activation. EPRV-9 is indeed more distantly related to BSGfV than is EPRV-7 (on average 0.74% versus 0.62% divergence, respectively) and has accumulated 35 more mutations. We next analyzed the mutations found in the three ORFs of the EPRVs (Table 5). Within the coding regions, EPRV-9 has accumulated 18 more substitutions than EPRV-7. EPRV-9 accumulated 11 nonsynonymous substitutions compared to the BSGfV genome, whereas EPRV-7 has only 2. Finally, we found three null mutations in ORFs of EPRV-9, but none in EPRV-7. We observed two substitutions leading to premature stop codons in ORF3 (fragments II and IV) and one adenosine insertion responsible for a frameshift leading to a stop codon in ORF1 (fragment Vc). The quality of the chromatograms of the EPRV-9 sequence confirmed that the three null substitutions exist in the cv. PKW genome and are not due to sequencing errors (data not shown).

TABLE 2. Sequencing ends of *M. balbisiana* cv. PKW BAC clones hybridizing with BSGfV probe

BAC clone	Probe	Primer	Sequence length (bp)	BLASTN		BLASTX	
				Result	E value	Result	E value
30-F18	BSGfV	Forward	519	*Oryza sativa* genomic DNA, chromosome 1	1e.09	WAK-like kinase (*Arabidopsis thaliana*)	2e.40
30-F18	BSGfV	Reverse	372	No significant similarity found		No significant similarity found	
64-H02	BSGfV	Forward	456	No significant similarity found		No significant similarity found	
64-H02	BSGfV	Reverse	458	No significant similarity found		No significant similarity found	
71-C19	BSGfV	Forward	644	No significant similarity found		No significant similarity found	
71-C19	BSGfV	Reverse	436	No significant similarity found		No significant similarity found	
96-J15	BSGfV	Forward	676	No significant similarity found		No significant similarity found	
96-J15	BSGfV	Reverse	346	No significant similarity found		No significant similarity found	
41-K09	BSGfV	Forward	922	No significant similarity found		No significant similarity found	
41-K09	BSGfV	Reverse	734	No significant similarity found		No significant similarity found	
48-D15	BSGfV	Forward	781	No significant similarity found		Hypothetical protein At2g28370 (*Arabidopsis thaliana*)	9e.07
48-D15	BSGfV	Reverse	NA[a]	NA	NA	NA	NA
72-M20	BSGfV	Forward	536	*Calycanthus fertilis* chloroplast genome	e.174	ATPase α subunit (*Calycanthus fertilis*)	8e.81
72-M20	BSGfV	Reverse	NA	NA	NA	NA	NA
73-C24	BSGfV	Forward	559	No significant similarity found		No significant similarity found	
73-C24	BSGfV	Reverse	480	*Calycanthus fertilis* chloroplast genome	1e.15	DNA-directed RNA polymerase common soapwort chloroplast	2e.05
94-116	BSGfV	Forward	576	No significant similarity found		No significant similarity found	
94-116	BSGfV	Reverse	449	*M. acuminata* retrotransposon monkey	8e.07	No significant similarity found	

[a] NA, not available.

Thus, we found strong conservation of all ORFs in each fragment constituting ERPV-7 and significant alterations of the ORFs in EPRV-9. EPRV-7 is therefore more likely than EPRV-9 to be involved in the restitution of infectious episomal BSGfV.

***Musa* genomic environment of BSGfV integrants in cv. PKW.** As demonstrated above, EPRV-7 and EPRV-9 are very similar in both structure and nucleotide sequence. This observation could be explained either by a duplication of an ancestral ERPV to another locus in the *Musa* genome or divergence of two allelic EPRVs at the same locus. We first annotated and aligned the two BAC clones MBP_71C19 and MBP_94116 (GenBank accession numbers AP009325 and AP009326, respectively). An 89.5-kb overlapping region between the two BAC clones with a very high sequence identity (99.7%) was found. The strongly conserved synteny of all genes in this overlapping area is shown in Table 6. These results are consistent with an allelic insertion of BSGfV in cv. PKW, where the two EPRVs are located on homologous chromosomes.

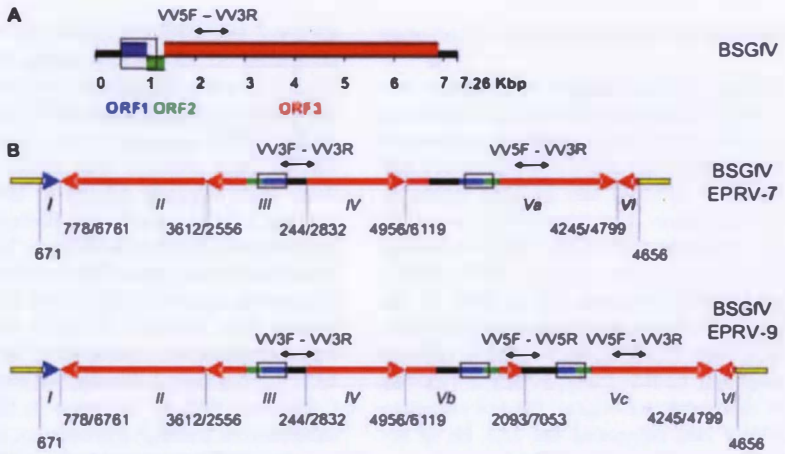

FIG. 3. BSGfV EPRV structures in cv. PKW. (A) Organization (linear view) of the BSGfV genome (GenBank AY3509). Blue, green, and red boxes indicate the three ORFs of the virus. The intergenic region is shown in black. (B) Structures of BSGfV EPRV-7 (top) and EPRV-9 (bottom) resulting from annotated BAC clones MBP_71C19 and MBP_94116, respectively. Arrows indicate the orientation of fragments of the BSGfV genome integrated in the *Musa* genome, shown in yellow. Blue, green, and red (the same code as used in panel A) refer to the different ORFs. Roman numerals identify the fragment. Numbers below each EPRV indicate the position of the fragment in the BSGfV genome. Open boxes indicate the region used in EPRV genotyping by PCR-RFLP DifGf F/R, and black arrows above the fragments indicate the regions amplified by multiplex-PCR with VV3F/R-VV5F/R.

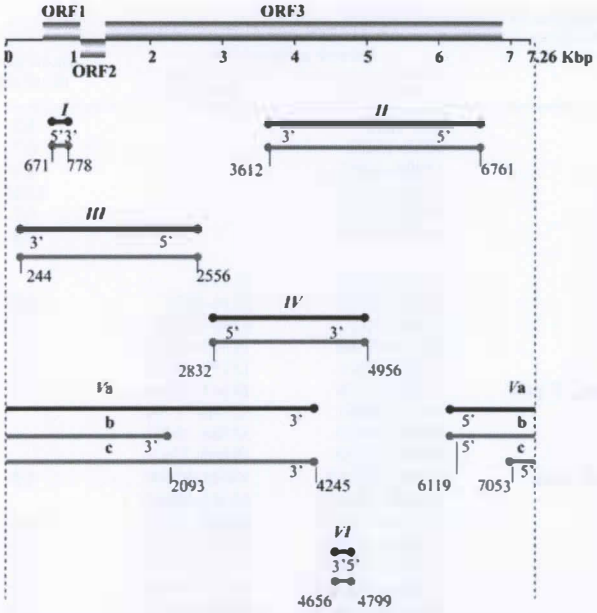

FIG. 4. Positions of EPRV fragments in the BSGfV genome. The genome of BSGfV is represented as in Fig. 3 (top). Lines below the BSGfV genome represent all of the fragments of EPRV-7 (in black) and EPRV-9 (in gray), and the circles indicate the boundaries of each fragment. Fragment names are indicated above each line. Fragments are arranged relative to their position in the BSGfV genome. The annotations 5′ and 3′ indicate the orientation of the fragments relative to the BSGfV genome.

Next, we analyzed the genomic environment of BSGfV integrations in cv. PKW. It was noted that BSGfV integrated into a gene-rich region of the *M. balbisiana* genome. Surprisingly, a finer annotation of the insertion locus revealed that both BSGfV EPRV-7 and EPRV-9 had integrated in the middle of a Ty3/gypsy retrotransposon (Fig. 5). Although the retroelement ORFs show signs of degradation, we found two long terminal repeats (LTRs; sequence TGTTAG-CTAACA) with a target site (sequence GT GGC) at each side, the primer-binding site (sequence TGGTAT CAGAC), and the polypurine track (sequence GAAGAGGAC GGG). The 3′ LTR (393 bp) is longer than the 5′ LTR (351 bp) because of a 42-bp insertion in the middle of the LTR. This 42-bp sequence is identical to the 5′ LTR sequence positions 144 to 176.

TABLE 3. Percentage of nucleotide identity within and between EPRVs[a]

EPRV	% Nucleotide identity				
	7 (FIII)	7 (FVa)	9 (FIII)	9 (FVb)	9 (FVc)
7 (FIII)	100				
7 (FVa)	**100**	100			
9 (FIII)	99.78	99.78	100		
9 (FVb)	99.78	99.78	**99.67**	100	
9 (FVc)	99.78	99.78	**99.78**	**99.67**	100

[a] Identity within EPRVs is indicated in boldface. A pairwise comparison was made on the same 1,847-bp sequence of BSGfV EPRV. The numbers 7 and 9 refer to EPRV-7 and EPRV-9, respectively. The specific EPRV fragment is indicated in parentheses.

TABLE 4. Number of mutations accumulated on each EPRV fragment compared to the known genome of BSGfV

Fragment	Size (bp)[a]	EPRV-7		EPRV-9	
		No. of mutations	Homology (%) with BSGfV	No. of mutations	Homology (%) with BSGfV
FI	101	0	100	0	100
FII	3,152	28	99.11	32	98.98
FIII	2,318	11	99.53	16	99.31
FIV	2,124	17	99.20	19	99.11
FVa	5,445 (5,383)	46	99.15		
FVb	3,290 (3,228)			38	98.82
FVc	4,456 (4,455)			32	99.28
FVI	141	1	99.29	1	98.99
Total		103		138	
Mean			99.38		99.26

[a] The numbers in parentheses refer to the size of the aligned sequences without gaps.

The retroelement itself is integrated in the fifth intron of a *mom* gene. The two split parts were called *mom* 3′ and 5′ part on the two *Musa* BAC clones. However, the N-terminal end of the corresponding protein is missing. The 5′ part of the *mom* gene might have been disrupted by another pseudogene or putative transposon found upstream.

Segregation of EPRV-7 and EPRV-9 in genetic crosses. To confirm the allelic EPRV insertion in cv. PKW, we monitored the segregation of BSGfV EPRV-7 and EPRV-9 carried by *M. balbisiana* cv. PKW (BB) (female parent) in the triploid (AAB) F1 progeny of a genetic cross with *M. acuminata* cv. IDN 110 4x (AAAA) (male parent). First, we confirmed the absence of BSGfV EPRV within the genomic DNA of *M. acuminata* cv. IDN 110 4x by PCR amplification using primers specific to ORF3 of BSGfV (Gf F/R) (Fig. 6). As expected, a product of the predicted size (476 bp) is obtained only from total genomic DNA of both cv. PKW (BB) and the entire F1 population. These results were confirmed by Southern blot hybridization. We also confirmed by using a Multiplex-Immuno-Capture PCR (29) that both parents were virus free (data not shown).

Next, we developed a set of specific molecular markers to genotype each EPRV. Since the two EPRVs are highly similar, we first developed a PCR-RFLP test that enables them to be distinguished due to a single-base substitution. The primer set

TABLE 5. Type of mutations accumulated in ORF1, ORF2, and ORF3 of both EPRVs relative to the ORFs of BSGfV genome

Type	No. of mutations			
	Shared between ERPV-7 and EPRV-9	Specific to EPRV-7	Specific to EPRV-9	Total
Synonymous substitution	32	1	7	40
Non synonymous substitution	15	2	11	28
Premature stop codon	0	0	2	2
Insertion leading to frameshift	0	0	1	1
Total	47	3	21	72

TABLE 6. Comparative analysis of annotated EPRVs[a]

Gene	Annotation	Putative gene position		No. of introns/ no. of exons
		MBP_71C19	MBP_94116	
I	Hypothetical protein	739–1212		1/2
II	Ribophorin I	4062–10471		6/7
III	Glycosyl transferase	11995–14863		2/3
IV	Epsin	17264–24154		12/13
VI	Putative lysine decarboxylase	29386–31885		6/7
VII	Pentatricopeptide (PPR) repeat protein	34768–37419		2/3
1	Ty3/gypsy-like retrotransposon, pseudogene	45714–52790	1–5536	
2	Hypothetical protein	55099–55341	8116–8358	1/2
3	Phenylalanine ammonia-lyase	55781–58032	8798–11049	1/2
4	Hypothetical protein	58333–59186	11349–12202	2/3
5	Zonadhesin	60143–67561	13159–20577	3/4
6	Transcriptional regulator related to the *mom* pseudogene, 3′ part	75953–78770	28967–31784	2/1
7	Ty3/gypsy retrotransposon, pseudogene, 3′ part	79531–80691	32546–33704	
EPRV	BSGfV integration	80691–93970	33704–49283	
8	Ty3/gypsy retrotransposon, pseudogene, 5′ part	93997–97330	49308–52643	
9	Transcriptional regulator related to *mom*, pseudogene, 5′ part	97779–100187	53095–55500	4/5
10	Putative transposon, pseudogene	101698–104653	57011–59852	
11	Hypothetical protein	107736–111174	63069–66507	7/8
12	Hypothetical protein	112684–117441	68017–72774	5/6
13	Hypothetical protein	118473–118806	73806–74139	2/3
14	Auxin-responsive protein	122614–123216	78063–78665	1/2
15	Actin-depolymerizing factor	125042–127808	80491–83257	1/2
16	Hypothetical protein	128615–129077	84064–84490	1/2
17	myb transcriptional factor		92254–93778	2/3
18	Putative wall-associated kinase		97501–100128	2/3
19	Hypothetical protein		102598–103133	2/3
20	Putative wall-associated kinase		105488–108091	2/3
21	Hypothetical protein		109438–110678	1/2
22	Putative wall-associated kinase		113308–115900	2/3

[a] The synteny between BAC clones MBP_71C19 (AP009325) and MBP_94116 (AP009326) is presented. Putative genes and their positions are given. The middle section of the table (separated by space above and below) represents the overlapping region between both BACs.

DifGf F/R amplifies the same fragment of 670 bp containing ORF1 and ORF2, which is present as two copies in EPRV-7 and three copies in EPRV-9 (Fig. 3B). The PCR products from EPRV-7 and EPRV-9 carry different numbers of sites for the restriction endonuclease TaaI (two versus one). Consequently, the PCR-RFLP test can distinguish between EPRV-7 and EPRV-9 and also indicates whether both EPRVs are present in the same genome, as is the case in cv. PKW (Fig. 7A). We subsequently screened the AAB F1 population and observed strict segregation of EPRV-7 and EPRV-9, which were found in 52.82 and 47.18% of the hybrid population, respectively. No heterozygote pattern was observed among the 142 hybrids tested. This result confirmed a monogenic 50-50 segregation ($df = 2, X^2 = 0.75, P = 0.80$) between the two BSGfV EPRVs.

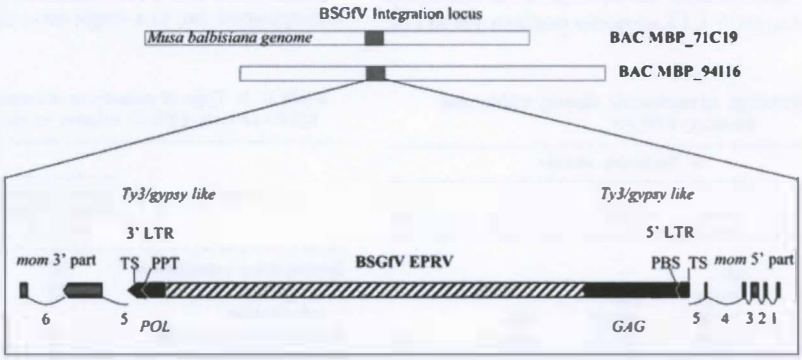

FIG. 5. *Musa* genomic environment of BSGfV EPRV. The orientation of the *mom* gene putative exons (gray arrows) and regions of the Ty3/gypsy-like retrotransposon (black arrows) are indicated. *mom* gene introns are numbered and indicated by thin lines. The 3′ and 5′ LTRs are indicated. TS, target site; PBS, primer binding site; PPT, polypurine tract. The regions of the GAG and POL polyproteins (Ty3/gypsy-encoded RT) in the retrotransposon are indicated.

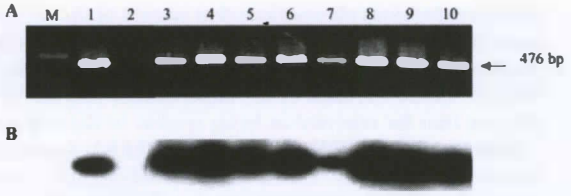

FIG. 6. PCR analysis of cv. PKW and cv. IDN 110 4x and their F1 progeny, using BSGfV specific primers. (A) Ethidium bromide-stained agarose gel analysis of PCR product (GfF [5′-ACGAACTATCACG ACTTGTTCAAGC-3′] and GfR [5′-TCGGTGGAATAGTCCTGAG TCTTC-3′]). (B) Southern blot hybridization of the gel shown in panel A using complete genome probes of BSGfV. Lane M, 1-kb ladder; lane 1, cv. PKW; lane 2, IDN 110 4x; lanes 3 to 6, F1 plants showing no sign of banana streak disease; lanes 7 to 10, F1 plants showing symptoms and BSV particles by immunosorbent electron microscopy as described by Lheureux et al. (30). Hybridization was performed according to the method of Sambrook et al. (45) using the two BSGfV probes (pCR-TOPO [1,262 bp] and pCR-TOPO [6,001 bp]).

In 2003, Lheureux et al. (30) demonstrated that a part of the progeny becomes infected de novo by BSV due to the activation of BSV EPRVs and that at least three different BSV species are expressed (data not shown), including the BSGfV studied here. Unfortunately, the PCR primer set DifGf F/R also recognizes the circular genome of BSGfV as a template for amplification. To avoid this cross-reaction, we developed a multiplex PCR using the primers VV3F/R and VV5F/R (Fig. 3B), which are highly specific for EPRV integrants, and able to differentiate EPRV-9 from EPRV-7. The VV5F/R primer set amplifies a 628-bp product with EPRV-9 only (Fig. 7B). The VV3F/R primer set amplifies a 376-bp fragment with both EPRVs and confirms EPRV amplification. In the multiplex PCR, the combination of primers VV5F and VV3R also amplifies a product of 1,012 bp from both EPRV and the circular BSGfV genome. We then screened the AAB population again and found exactly the same results as with the PCR-RFLP test.

Finally, in order to detect the possible recombination of BSGfV EPRVs in the hybrid progeny, we designed two additional PCR markers surrounding each of the two integration sites. The Spe7F/R primer set (Fig. 7C) amplifies a region located 28.2 kb upstream of EPRV-7, and the Spe9bisF/R primer set amplifies a region located 26.8 kb downstream of EPRV-9. None of the progeny showed a recombinant profile with either no amplification or both PCR products in the same individual. This latter genotyping method further confirmed the strict segregation of EPRV-7 and -9 in the progeny. Thus, three experimental approaches independently confirmed that the two BSGfV EPRVs, EPRV-7 and EPRV-9, are located on homologous chromosomes in the genome of *M. balbisiana* cv. PKW. We conclude that EPRV-7 and EPRV-9 are two alleles of the same locus in cv. PKW.

Which of the two EPRVs, EPRV-7 or EPRV-9, is infectious? To demonstrate the infectious nature of EPRVs and determine which allele—EPRV-7 and EPRV-9—is infectious, we genotyped the BSGfV particles expressed in the AAB progeny. We developed an IC-multiplex PCR-RFLP method to genotype the molecular EPRV signature of BSGfV particles (Fig. 8). The IC step allows the capture of viral particles by a BSV polyclonal antiserum. Then, a single multiplex PCR specifically amplifies a 670-bp product from immunocaptured BSGfV with the DifGfF/R primers, whereas the primer set Act1F/R amplifying a 420-bp product from *Musa* housekeeping actin gene

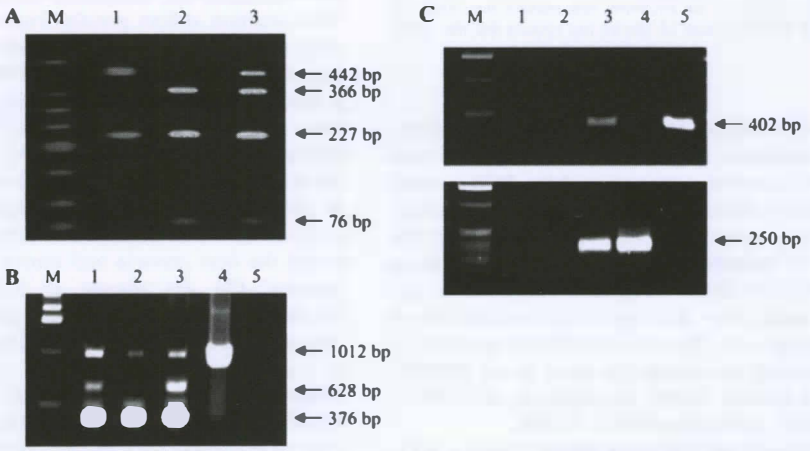

FIG. 7. Genotyping of BSGfV EPRV-7 and EPRV-9 (A, B, and C) and detection of recombinant EPRV (C). (A) PCR DifGf F/R-RFLP to genotype BSGfV EPRVs in cv. PKW. Endonuclease TaaI discriminates between EPRV-7 and EPRV-9; amplification products carry two versus one restriction sites, respectively. Lane M, DNA ladder (low molecular weight; Invitrogen). Digestion of PCR product DifGfF/R on clone MBP_94116 carrying EPRV-9 (lane 1), on clone MBP_71C19 carrying EPRV-7 (lane 2), and on *M. balbisiana* cv. PKW carrying both EPRVs (lane 3) was performed. No amplification product was seen on *M. acuminata* cv. IDN 110 4x (data not shown). (B) Multiplex PCR with primers VV3 and VV5 for BSGfV EPRV genotyping. Primers VV3F/R amplify a 376-bp product in both EPRVs, primers VV5F/R amplify a 628-bp product in EPRV-9 only, and primers VV5F/R amplify a 1,012-bp product on both EPRVs and the BSGfV circular genome. Lane M, 1-kb ladder (Invitrogen). Lane 1, BAC MBP_94116; lane 2, BAC MBP_71C19; lane 3, DNA of *M. balbisiana* cv. PKW; lane 4, DNA of *M. acuminata* infected by BSGfV; lane 5, PCR negative control. (C) PCR detection of recombination between the two BSGfV EPRVs. PCR results with Spe7F/R (top) specific to EPRV-7 and Spe9bisF/R (bottom) specific to EPRV-9. Lane M, 1-kb ladder (Invitrogen). Lane 1, negative PCR control (water); lane 2, *M. acuminata* cv. IDN 110 4x genomic DNA; lane 3, *M. balbisiana* cv. PKW genomic DNA; lane 4, BAC MBP_94116; lane 5, BAC MBP_71C19.

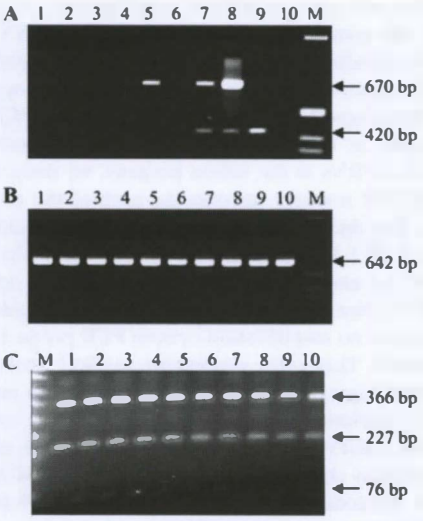

FIG. 8. Genotyping of BSGfV viral particles in infected hybrids. (A) IC-multiplex PCR allows specific detection of BSGfV particles (DifGfF/R, 670-bp product) and a monitoring of plant DNA contaminations (ActinIF/R, 420-bp product). Lane M, 1-kb ladder (Invitrogen); lanes 1 to 6, coated plant extracts; lanes 7 to 10, plant total DNA. Lanes 1 to 3 show results for F1 AAB hybrids; lanes 4 to 6 show results for the IC control (lane 4, *M. balbisiana* cv. PKW; lane 5, *M. acuminata* cv. Grande Naine infected by BSGfV; lane 6, *M. acuminata* cv. Grande Naine BSGfV-free). Lanes 7 to 10 show results for the PCR control (lane 7, *M. balbisiana* cv. PKW; lane 8, *M. acuminata* cv. Grande Naine infected by BSGfV; lane 9, *M. acuminata* cv. Grande Naine BSGfV-free; and lane 10, water control). (B) Nested PCR using the internal primers VV1F/GfM2 (642-bp product) and increasing PCR product quantity from diluted DifGfF/R PCR product of infected hybrids. Lanes 1 to 10 show the results for AAB F1 hybrids infected with BSGfV; lane M, 1-kb ladder (Invitrogen). (C) TaaI RFLP test (described in Fig. 7A) to genotype the molecular EPRV signature of the viral BSGfV particle. Lanes 1 to 10 show the results for AAB F1 hybrids infected with BSGfV; lane M shows the results for the 50-bp ladder (NEB).

monitors the possible residual *Musa* genomic DNA containing BSV EPRV contaminations (Fig. 8A). A final nested PCR with internal primers increases the quantities of the PCR product (Fig. 8B), allowing an efficient digestion by TaaI endonuclease (Fig. 8C) to a final genotyping of BSGfV viral particles. We screened the 166 F1 hybrids by using this method. Seventeen hybrids were infected by BSGfV (Fig. 8A). There was no amplification of the actin gene, attesting to the amplification of episomal viral genome only. The molecular EPRV signature of viral particles recorded was always the same as for EPRV-7 (Fig. 8C); no viral particle carried the signature of EPRV-9. All 17 infected plants harbor the EPRV-7 allele.

These results demonstrate that allele EPRV-7 only is infectious and is able to release infectious BSGfV, causing systemic infection.

DISCUSSION

Using a high-resolution hybridization method, we demonstrated that BSGfV sequences are integrated only in the genome of *M. balbisiana* cultivar cv. PKW and are absent from two other common cultivars of *M. acuminata* tested. This result reinforces previous observations that species of the BSV clade sensu stricto, to which BSGfV belongs, are integrated mainly in the B genome (15, 16), whereas a minority of BSV species, e.g., BSCavV (Iskra-Caruana et al., unpublished) and BSAcVNV (31) are thus far reported as being specific to the A genome.

Only two BSGfV EPRVs, EPRV-7 and EPRV-9, exist in the nuclear genome of cv. PKW, and their integration is unique among the EPRVs described thus far. First, despite the fact that the viral genome appears fragmented, inverted, and partially repeated, surprisingly each EPRV contains the full-length genome of BSGfV. The EPRVs also each contain all of the genetic information needed for "reconstruction" of a functional BSGfV genome very similar to that of the infectious BSGfV virus, with EPRV-7 being the most conserved and showing no evidence of ORF degradation. In the progeny, all hybrids infected with BSGfV harbor EPRV-7, and all BSGfV particles in these hybrids showed an EPRV-7 signature. We therefore demonstrate that EPRV-7 is the infectious EPRV in our pathosystem. Furthermore, EPRV-7 and EPRV-9 are highly similar in general structure and nucleotide sequence and share a common surrounding genomic environment, as detected by contig building from BAC fingerprints, as well as sequencing of BACs carrying the two types of BSGfV EPRVs. This situation either could be due to duplication of an ancestral EPRV in a different locus of the genome or could have originated from a divergence of two EPRVs located on homologous chromosomes. By examining EPRV segregation in interspecific crosses, we demonstrated that EPRV-7 and EPRV-9 are two alleles of the same locus. This integration is therefore the consequence of a single integration event with no subsequent copy number increase. Not only do the two alleles share great similarity of sequence and structure but also only a few mutations differentiate them from the BSGfV genome, thus indicating the integration event to be relatively recent. This situation differs greatly from other previously studied cases of EPRVs. First, the only described integration of infectious EPRV is one of the many ePVCV in the *Petunia hybrida* genome. This integrant is a tandem direct repeat (i.e., in the same orientation) of the full PVCV genome (41). Second, EPRV sequences of PVCV, TVCV, and several BSV-like species found in petunia, tobacco, and banana (*M. acuminata* and *M. balbisiana*), respectively, are highly decayed and are found as numerous small fragments of badnaviral genome, dispersed within the host genome and usually referred to as "dead sequences" (20, 27). Finally, all EPRVs described thus far, whether they contain the full viral genome or not, reach a high copy number in their host genome through a dynamic process of accumulation and elimination (17). EPRV copy number ranges from dozens to several hundreds—as, for example, with BSOIV EPRV in the cultivar Obino l'Ewai (AAB) (21), ePVCV in petunia (41), or LycEPRV in tomato (52)—to thousands for NsEPRV in tobacco (37) and NtoEPRV in *Nicotiana tomentosiformis* (17).

Endogenous viral sequences are a common constituent of many plant genomes (53). Integration generally results from an active mechanism, e.g., retroviral integrases, but this does not apply to pararetroviruses. Indeed, despite the fact that the petunia vein clearing pararetrovirus polyprotein contains two motifs resembling the catalytic domain motifs of integrase (42), no further sequence homology to putative integrase domains

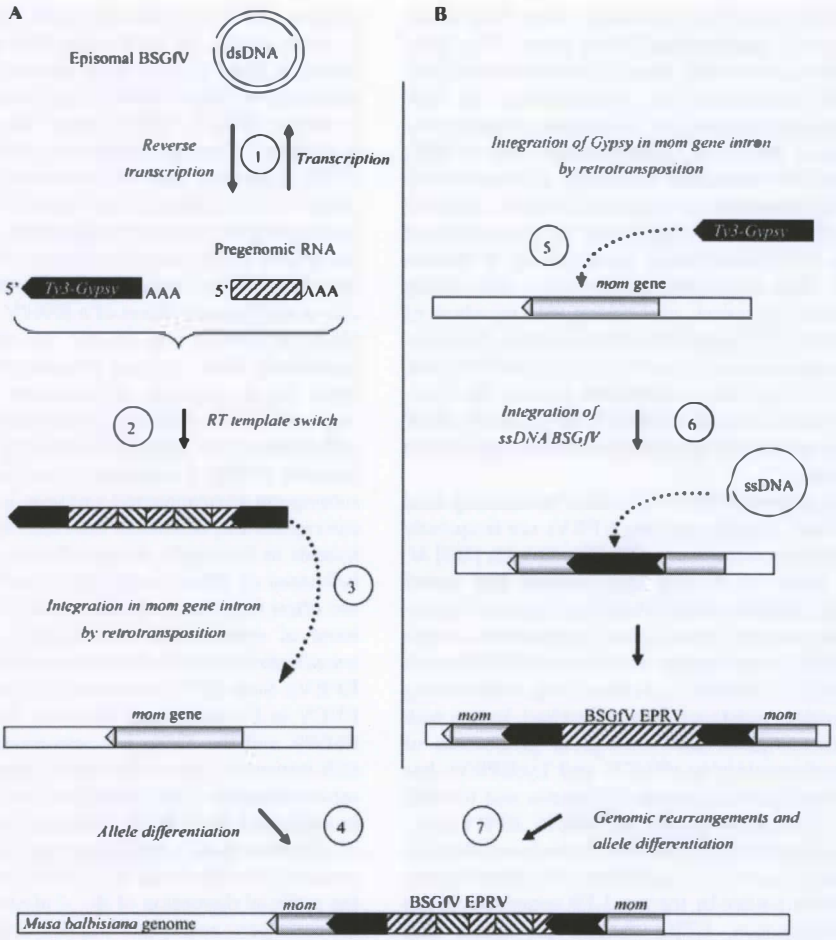

FIG. 9. Scenarios for BSGfV integration in the *M. balbisiana* nuclear genome. (A) Hypothesis 1: use of Ty3/gypsy retroelement for BSGfV integration. Viral RT leads to a pregenomic viral RNA (step 1) during episomal BSGfV infection. A template switch of retrotransposon RT between pregenomic viral RNA and a replicating Ty3/gypsy RNA may form a chimeric RNA (step 2). This step leads to the rearrangements of BSGfV EPRV: fragmentation, inversion, and duplication. Complete retrotransposition of the Ty3/gypsy element leads to its integration (step 3) into the genome of *M. balbisiana* (within intron 5 of the *mom* gene). Subsequent genomic change such as recombination may lead to the structural differences observed between the two alleles 7 and 9. (B) Hypothesis 2: Ty3/gypsy retroelement integrated in the fifth intron of the *mom* gene (step 5) during retrotransposition. Double-strand break repair could account for illegitimate recombination with the single-stranded DNA template generated during BSGfV reverse transcription and integration of the BSGfV genome (step 6). A subsequent genomic change, e.g., recombination, may have lead to both the fragmentation, inversion, and duplication observed in BSGfV EPRV and the structural differences between the alleles 7 and 9 (step 7).

of retroelements could be found (20), and no experimental data confirm this function. Instead, plant pararetroviruses are thought to integrate in the host genome via accidental illegitimate recombination during the minichromosome phase. We propose two scenarios to explain the integration process and the final EPRV structure observed, taking into account both the BSGfV insertion locus in *Musa* chromosomes and the complex structure of EPRVs. One possible scenario (Fig. 9A) assumes recombination at the RNA level between the pregenomic viral RNA resulting from BSGfV infection and the RNA of a retrotranscribing Ty3/gypsy retrotransposon existing in the *Musa* genome. RNA recombination may originate from a template switch (5, 55) by the Ty3/gypsy reverse transcriptase (RT). An RT template switch between several chimeric pregenomic RNAs could also explain the rearrangements of viral sequences and thus the complex EPRV structure. If the chimeric RNA molecule produced retained its ability to fulfill the retrotranscription process, it could have integrated into the host genome to form the BSGfV integration observed today. In our model, integration of a chimeric Ty3/gypsy-BSGfV transposable element occurred in the fifth intron of the *mom* gene. Transposition of retrotransposons in gene introns is a frequent phenomenon observed, for example, in mammalian (47) and rice genomes (59). A second possible scenario (Fig. 9B) proposes integration of the Ty3/gypsy retrotransposon into the *mom* gene intron as a first event, predating integration of BSGfV DNA into the Ty3/gypsy element itself. It is generally acknowledged that integration of viral DNA occurs in the

nucleus during viral replication and results from illegitimate recombination after a double-strand break repair. The presence of gaps in the open circular form of pararetroviral DNA may facilitate this mechanism (26). Furthermore, the Ty3/gypsy retrotransposon belongs to the *Metaviridae*, a family phylogenetically close to the family *Caulimoviridae* (19). In 2005, Puchta (40) showed that sequence homology, or microhomology, enhances recombination, in this case between badnaviruses and the retrotransposon responsible for integration of the viral genome. EPRVs are found preferentially in heterochromatin rather than euchromatin (52), often colocalizing with retrotransposon sequences, particularly with members of the family *Metaviridae* (Ty3/gypsy retrotransposons). This general feature was also observed for ePVCV (41), NsEPRV (26), NtoEPRV (17), LycEPRV (52), and BSOIV EPRV (38). However, the exact genomic location of BSGfV EPRV in cv. PKW remains unknown, and in situ hybridization will be required to answer this question.

EPRVs are also suspected to be beneficial by inducing viral resistance in the host. Species carrying EPRVs are frequently resistant to the corresponding virus (TVCV, diploids [BB] *M. balbisiana*), and Maori et al. (36) hypothesized that *Israeli acute paralysis virus* (dicistrovirus) integration into the honeybee (*Apis mellifera*) genome could explain bee resistance to this virus. This hypothesis could explain why infectious EPRVs are maintained by natural selection in plants as long as they bring a homology-dependent resistance (gene silencing). In line with this, moderate transcription and subsequent production of small RNAs complementary to ePVCV and LycEPRVs has recently been proved in the genomes of petunia and tomato, respectively (39, 52). Unfortunately, no BSOIV EPRV transcription or small interfering RNAs have been found thus far in cv. PKW (Iskra-Caruana et al., unpublished). Nevertheless, BSGfV EPRV is surrounded by the two LTR sequences of the Ty3/gypsy retrotransposon. LTRs contain promoters that might facilitate the expression of BSGfV EPRV. The RNA transcript from EPRV might undergo subsequent recombination, for instance, using endogenous *Musa* RT, thereby becoming pathogenic. In this respect, further studies on BSV EPRV expression and the levels of methylation will be required.

Among parasites, EPRVs are unusual pathogens. Each partner interacts at the genetic and genomic level and is engaged in an arms race. Probably in response to their potential harmful effects, natural selection has favored several host defenses against EPRV activation. First, the fragmentation, duplication, and inversion of EPRV sequences potentially decrease the probability that an EPRV can induce the production of a functional and infectious BSGfV genome. Maintaining such disorganization of integrated BSGfV genomes could be an evolving situation of host protection to hamper EPRV activation. Second, DNA and histone methylation are thought to explain the transcriptional silencing observed in ePVCV and LycEPRVs (39, 52). Although cv. PKW appears resistant to both EPRV expression and BSV infection despite harboring infectious EPRV, regulation of EPRV expression by DNA methylation has not been demonstrated, at least for BSOIV in cv. PKW (Iskra-Caruana et al., unpublished). We assume from our results that BSGfV integration in cv. PKW is a recent event from an evolutionary point of view. BSV integration is perhaps too recent for a resistance to BSV based on RNA interference-mediated silencing from expressed EPRVs, like that observed in other plants, to have evolved in *Musa* plants. From the pathogen point of view, three factors might be linked with the activation of BSGfV EPRV. First, because BSGfV integration is recent, BSGfV EPRVs have not yet evolved into "dead sequences." The few mutations accumulated within BSGfV EPRV sequences were not numerous enough to result in the decay of viral ORFs in the case of EPRV-7. It is generally acknowledged referring to hypothesis developed from the partial BSOIV EPRV described in the AAB cv. Obino l'Ewai (38) that homologous recombination in the plant genome plays a role in the reconstruction of a BSGfV genome from functional ORFs in EPRVs (13, 24, 46), but the link is not yet firmly established. Next, a strong activation of retroelement transposition due to a release of epigenetic silencing is observed in response to UV exposure, temperature, radiation, wounding, cell culture, and polyploidization (4, 48). *Musa* hybrids are triploids (AAB), propagated by in vitro culture, and undergo subsequent environmental variation in the field. These stresses can explain why activation was restricted strictly to interspecific hybrids in our study, despite the fact that the genome of *M. balbisiana* cv. PKW carries infectious EPRVs. Because EPRVs are often found near or embedded in *Metaviridae* elements, a burst of retroelement transposition might facilitate EPRV transcription and therefore the activation of infectious BSGfV EPRVs. Such EPRV activation in hybrids is also observed for PVCV in *P. hybrida* and TVCV in *Nicotiana tabacum*. Lastly, BSGfV and the Ty3/gypsy retrotransposon are found in the fifth intron of a *mom Musa* gene. Integration of BSGfV and a retrotransposon in the *mom* gene intron might have disturbed its expression in cv. PKW. This could result in a loss of function of the *mom* gene, explaining why it subsequently became a pseudogene with decay in its coding sequence. Astonishingly, the artificial disruption of the *Arabidopsis thaliana* ortholog of the *mom* gene reactivates the transcription of previously silent genes (2) and repetitive sequences (57). It is therefore tempting to speculate that, as in *A. thaliana*, *mom* gene disruption by BSGfV EPRV and the Ty3/gypsy retrotransposon reconstitutes the expression of previously repressed genes. *mom* gene disruption might facilitate the expression of BSGfV EPRV itself, but also other BSV EPRVs present in the cv. PKW genome, thereby increasing the probability of their activation.

BSV sequences found integrated in the genome of the host banana (genus *Musa*) are of great concern since several BSV species integrated in *M. balbisiana* are infectious. Here, we report for the first time the full molecular organization and functional analysis of one such sequence present in the genome of the diploid *M. balbisiana* cv. PKW (BB). This viral sequence corresponds to the BSV species Goldfinger infectious in interspecific hybrids obtained by genetic crosses involving cv. PKW. Knowledge of the molecular organization of BSV EPRVs in the *Musa* genome is of crucial interest to researchers and plant breeders in order to overcome problems caused by their presence in banana plant genomes. Actually, the main difficulty comes from the fact that cv. PKW and *M. balbisiana* genotypes in general harbor at least three other integrated BSV species—BSOIV, BSImV, and *Banana streak Mysore virus*, each of them with several EPRVs—and that BSOIV and BSImV EPRVs are also infectious (Iskra-Caruana, unpublished). Identifying genetic resources free from BSV EPRVs and producing recom-

binant *Musa* genotypes having lost the set of infectious EPRV corresponding to the three BSV species are the challenges currently facing both scientists and breeders. Details of EPRV activation processes, including recombination at the plant DNA level and viral and host factors involved in the production of infectious BSGfV genomes in hybrids, need to be further characterized.

ACKNOWLEDGMENTS

Construction of the BAC libraries was supported by CIRAD, Academy of Sciences of the Czech Republic (project A6038201), the International Atomic Energy Agency (research contract 12230/RBF), and the French Ministry of Foreign Affairs (COCOP project 17/01). The study was undertaken as a part of the Global Programme for Musa Improvement (PROMUSA) coordinated by BIOVERSITY (previously named INIBAP). P.G. is supported by a CIRAD/Région Languedoc-Roussillon Ph.D. grant.

We are very grateful to Liying Zhang for providing the two clones of BSGfV; Franc-Christophe Baurens and Stéphanie Sidibe Boes for help with the Image and FPC softwares; and Kozue Kamiya, Hiroyuki Kanamori, and Takuji Sasaki for performing the sequencing of the two BAC clones (MBP_71C19 and MBP_94116).

REFERENCES

1. **Altschul, S., W. Gish, W. Miller, E. Myers, and D. Lipman.** 1990. Basic local alignment search tool. J. Mol. Biol. 215:403–410.
2. **Amedeo, P., Y. Habu, K. Afsar, O. M. Scheid, and J. Paszkowski.** 2000. Disruption of the plant gene MOM releases transcriptional silencing of methylated genes. Nature 405:203–206.
3. **Berriman, M., and K. Rutherford.** 2003. Viewing and annotating sequence data with Artemis. Brief Bioinform. 4:124–132.
4. **Capy, P., G. Gasperi, C. Biemont, and C. Bazin.** 2000. Stress and transposable elements: co-evolution or useful parasites? Heredity 85:101–106.
5. **Cocquet, J., A. Chong, G. L. Zhang, and R. A. Veitia.** 2006. Reverse transcriptase template switching and false alternative transcripts. Genomics 88: 127–131.
6. **Dahal, G., D. A. Hughes, G. Thottappilly, and B. E. L. Lockhart.** 1998. Effect of temperature on symptom expression and expression and reliability of banana streak badnavirus detection in naturally infected plantain and banana (*Musa* spp). Plant Dis. 82:16–21.
7. **Dallot, S., P. Acuna, C. Rivera, P. Ramirez, F. Cote, B. E. L. Lockhart, and M. L. Caruana.** 2001. Evidence that the proliferation stage of micropropagation procedure is determinant in the expression of *Banana streak virus* integrated into the genome of the FHIA 21 hybrid (*Musa* AAAB). Arch. Virol. 146:2179–2190.
8. **Ewing, B., and P. Green.** 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res. 8:186–194.
9. **Ewing, B., L. Hillier, M. C. Wendl, and P. Green.** 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome. Res. 8:175–185.
10. **Fargette, D., G. Konate, C. Fauquet, E. Muller, M. Peterschmitt, and J. M. Thresh.** 2006. Molecular ecology and emergence of tropical plant viruses. Ann. Rev. Phytopathol. 44:235–260.
11. **Fauquet, C. M., M. A. Mayo, J. Maniloff, U. Desselberger, and L. A. Ball.** 2005. Virus taxonomy: Eighth report of the International Committee on Taxonomy of Viruses. Elsevier/Academic Press, Inc., New York, NY.
12. **Folliot, M., S. Galzi, N. Laboureau, M.-L. Caruana, P.-Y. Teycheney, and F.-X. Côte.** 2005. Risk assessment of spreading *Banana Streak Virus* (BSV) through *in vitro* culture. XIIIth International Congress of Virology, San Francisco, CA.
13. **Gaut, B. S., S. I. Wright, C. Rizzon, J. Dvorak, and L. K. Anderson.** 2007. Recombination: an underappreciated factor in the evolution of plant genomes. Nat. Rev. Genet. 8:77–84.
14. **Gawel, N. J., and R. L. Jarret.** 1991. A modified CTAB DNA extraction procedure for *Musa* and *Ipomea*. Plant Mol. Biol. Reporter 9:262–266.
15. **Geering, A. D. W., N. E. Olszewski, G. Dahal, J. E. Thomas, and B. E. L. Lockhart.** 2001. Analysis of the distribution and structure of integrated *Banana streak virus* DNA in a range of *Musa* cultivars. Mol. Plant Pathol. 2:207–213.
16. **Geering, A. D. W., N. E. Olszewski, G. Harper, B. E. L. Lockhart, R. Hull, and J. E. Thomas.** 2005. Banana contains a diverse array of endogenous badnaviruses. J. Gen. Virol. 86:511–520.
17. **Gregor, W., M. F. Mette, C. Staginnus, M. A. Matzke, and A. J. M. Matzke.** 2004. A distinct endogenous pararetrovirus family in *Nicotiana tomentosiformis*, a diploid progenitor of polyploid tobacco. Plant Physiol. 134:1191–1199.
18. **Hall, T. A.** 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symp. Ser. 41:95–98.
19. **Hansen, C., and J. S. Heslop-Harrison.** 2004. Sequences and phylogenies of plant pararetroviruses, viruses, and transposable elements. Adv. Bot. Res. Adv. Plant Pathol. 41:165–193.
20. **Harper, G., R. Hull, B. Lockhart, and N. Olszewski.** 2002. Viral sequences integrated into plant genomes. Annu. Rev. Phytopathol. 40:119–136.
21. **Harper, G., J. O. Osuji, J. S. P. Heslop-Harrison, and R. Hull.** 1999. Integration of *Banana streak badnavirus* into the *Musa* genome: molecular and cytogenetic evidence. Virology 255:207–213.
22. **Hull, R.** 1999. Classification of reverse transcribing elements: a discussion document. Arch. Virol. 144:209–214.
23. **Hull, R., and S. N. Covey.** 1995. Retroelements: propagation and adaptation. Virus Genes 11:105–118.
24. **Hull, R., G. Harper, and B. Lockhart.** 2000. Viral sequences integrated into plant genomes. Trends Plant Sci. 5:362–365.
25. **Iskra Caruana, M. L., F. Lheureux, J. C. Noa-Carrazana, P. Piffanelli, F. Carreel, C. Jenny, N. Laboureau, and B. E. L. Lockhart.** 2003. Unstable balance of relation between pararetrovirus and its host plant: the BSV-EPRV banana pathosystem. p. 8. EMBO Workshop: Genomic Approaches in Plant Virology, Keszthely, Hungary.
26. **Jakowitsch, J., M. F. Mette, J. van der Winden, M. A. Matzke, and A. J. M. Matzke.** 1999. Integrated pararetroviral sequences define a unique class of dispersed repetitive DNA in plants. Proc. Natl. Acad. Sci. USA 96:13241–13246.
27. **Kunii, M., M. Kanda, H. Nagano, I. Uyeda, Y. Kishima, and Y. Sano.** 2004. Reconstruction of putative DNA virus from endogenous rice tungro bacilliform virus-like sequences in the rice genome: implications for integration and evolution. BMC Genomics 5:80.
28. **Kurtz, S., and C. Schleiermacher.** 1999. REPuter: fast computation of maximal repeats in complete genomes. Bioinformatics 15:426–427.
29. **Le Provost, G., M. L. Iskra-Caruana, I. Acina, and P. Y. Teycheney.** 2006. Improved detection of episomal banana streak viruses by multiplex immunocapture PCR. J. Virol. Methods 137:7–13.
30. **Lheureux, F., F. Carreel, C. Jenny, B. Lockhart, and M. Iskra-Caruana.** 2003. Identification of genetic markers linked to banana streak disease expression in inter-specific *Musa* hybrids. TAG Theor. Appl. Genet. 106: 594–598.
31. **Lheureux, F., N. Laboureau, E. Muller, B. E. Lockhart, and M. L. Iskra-Caruana.** 2007. Molecular characterization of banana streak acuminata Vietnam virus isolated from *Musa acuminata siamea* (banana cultivar). Arch. Virol. 152:1409–1416.
32. **Lheureux, F.** 2002. Etude des mécanismes génétiques impliqués dans l'expression des séquences EPRVs pathogènes des bananiers au cours de croisements génétiques interspécifiques. Ph.D. thesis. Université Sciences et Techniques du Languedoc, Montpellier, France.
33. **Lockhart, B., and D. Jones.** 2000. Banana streak, p. 263–274. *In* D. R. Jones (ed.), Diseases of banana, abaca, and enset. CAB International, Wallingford, United Kingdom.
34. **Lockhart, B. E., J. Menke, G. Dahal, and N. E. Olszewski.** 2000. Characterization and genomic analysis of tobacco vein clearing virus, a plant pararetrovirus that is transmitted vertically and related to sequences integrated in the host genome. J. Gen. Virol. 81:1579–1585.
35. **Lukashin, A. V., and M. Borodovsky.** 1998. GeneMark.hmm: new solutions for gene finding. Nucleic Acids Res. 26:1107–1115.
36. **Maori, E., E. Tanne, and I. Sela.** 2007. Reciprocal sequence exchange between non-retroviruses and hosts leading to the appearance of new host phenotypes. Virology 362:342–349.
37. **Mette, M. F., T. Kanno, W. Aufsatz, J. Jakowitsch, J. van der Winden, M. A. Matzke, and A. J. M. Matzke.** 2002. Endogenous viral sequences and their potential contribution to heritable virus resistance in plants. EMBO J. 21: 461–469.
38. **Ndowora, T., G. Dahal, D. LaFleur, G. Harper, R. Hull, N. E. Olszewski, and B. Lockhart.** 1999. Evidence that badnavirus infection in *Musa* can originate from integrated pararetroviral sequences. Virology 255:214–220.
39. **Noreen, F., R. Akbergenov, T. Hohn, and K. R. Richert-Poggeler.** 2007. Distinct expression of endogenous petunia vein clearing virus and the DNA transposon dTph1 in two *Petunia hybrida* lines is correlated with differences in histone modification and siRNA production. Plant J. 50:219–229.
40. **Puchta, H.** 2005. The repair of double-strand breaks in plants: mechanisms and consequences for genome evolution. J. Exp. Bot. 56:1–14.
41. **Richert-Poggeler, K. R., F. Noreen, T. Schwarzacher, G. Harper, and T. Hohn.** 2003. Induction of infectious petunia vein clearing (pararetro) virus from endogenous provirus in petunia. EMBO J. 22:4836–4845.
42. **Richert-Poggeler, K. R., and R. J. Shepherd.** 1997. Petunia vein-clearing virus: a plant pararetrovirus with the core sequences for an integrase function. Virology 236:137–146.
43. **Safar, J., J. C. Noa-Carrazana, J. Vrana, J. Bartos, O. Alkhimova, X. Sabau, H. Simkova, F. Lheureux, M. L. Caruana, J. Dolezel, and P. Piffanelli.** 2004. Creation of a BAC resource to study the structure and evolution of the banana (*Musa balbisiana*) genome. Genome 47:1182–1191.

44. **Salamov, A. A., and V. V. Solovyev.** 2000. Ab initio gene finding in *Drosophila* genomic DNA. Genome Res. 10:516–522.
45. **Sambrook, J., and D. W. Russell.** 2001. Molecular cloning: a laboratory manual, 3rd ed. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
46. **Schuermann, D., J. Molinier, O. Fritsch, and B. Hohn.** 2005. The dual nature of homologous recombination in plants. Trends Genet. 21:172–181.
47. **Sironi, M., G. Menozzi, G. P. Comi, M. Cereda, R. Cagliani, N. Bresolin, and U. Pozzoli.** 2006. Gene function and expression level influence the insertion/fixation dynamics of distinct transposon families in mammalian introns. Genome Biol. 7:R120.
48. **Slotkin, R. K., and R. Martienssen.** 2007. Transposable elements and the epigenetic regulation of the genome. Nat. Rev. Genet. 8:272–285.
49. **Soderlund, C., S. Humphray, A. Dunham, and L. French.** 2000. Contigs built with fingerprints, markers, and FPC V4.7. Genome Res. 10:1772–1787.
50. **Solovyev, V., A. Zharkikh, and N. Kolchanov.** 1985. Context analysis of polynucleotide sequences: methods of detecting non-random repeats. I. Direct repeats in genes of β, β′, σ subunits of *Escherichia coli* RNA-polymerase. Mol. Biol. 19:524–536. (In Russian.)
51. **Sonnhammer, E. L., and R. Durbin.** 1995. A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. Gene 167:GC1–GC10.
52. **Staginnus, C., W. Gregor, M. F. Mette, C. H. Teo, E. G. Borrobo-Fernandez, M. L. Machado, M. Matzke, and T. Schwarzacher.** 2007. Endogenous pararetroviral sequences in tomato (*Solanum lycopersicum*) and related species. BMC Plant Biol. 7:24.
53. **Staginnus, C., and K. R. Richert-Poggeler.** 2006. Endogenous pararetroviruses: two-faced travelers in the plant genome. Trends Plant Sci. 11:485–491.
54. **Suiston, J., F. Mallett, R. Durbin, and T. Horsnell.** 1989. Image analysis of restriction enzyme fingerprint autoradiograms. Comput. Appl. Biosci. 5:101–106.
55. **Temin, H. M.** 1993. Retrovirus variation and reverse transcription: abnormal strand transfers result in retrovirus genetic variation. Proc. Natl. Acad. Sci. USA 90:6900–6903.
56. **Thompson, J. D., D. G. Higgins, and T. J. Gibson.** 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties, and weight matrix choice. Nucleic Acids Res. 22:4673–4680.
57. **Vaillant, I., I. Schubert, S. Tourmente, and O. Mathieu.** 2006. MOM1 mediates DNA-methylation-independent silencing of repetitive sequences in *Arabidopsis*. EMBO Rep. 7:1273–1278.
58. **Vilarinhos, A. D., P. Piffanelli, P. Lagoda, S. Thibivilliers, X. Sabau, F. Carreel, and A. D'Hont.** 2003. Construction and characterization of a bacterial artificial chromosome library of banana (*Musa acuminata* Colla). Theor. Appl. Genet. 106:1102–1106.
59. **Wang, G. D., P. F. Tian, Z. K. Cheng, G. Wu, J. M. Jiang, D. B. Li, Q. Li, and Z. H. He.** 2003. Genomic characterization of Rim2/Hipa elements reveals a CACTA-like transposon superfamily with unique features in the rice genome. Mol. Genet. Genomics 270:234–242.

2 La recombinaison homologue : un mécanisme d'activation de l'EPRV infectieux du *Banana streak GF virus* chez le bananier *Musa balbisiana* cv. PKW

2.1 Objectifs généraux

L'étude précédente concernant la caractérisation de l'EPRV infectieux du BSGFV chez le cv. PKW a révélé que le génome viral dans l'EPRV n'est pas constitué d'un fragment unique, mais de plusieurs fragments réarrangés.

L'objectif de cette étude est de proposer un modèle théorique d'activation des EPRV reposant sur des évènements de recombinaison homologue, afin d'expliquer la restitution du génome viral fonctionnel circulaire à partir d'un EPRV linéaire et réarrangé.

Le modèle proposé a ensuite été validé, en nous assurant d'une part que les régions de l'EPRV impliquées dans la formation du génome circulaire sont indemnes de mutations délétères, et d'autre part, en détectant expérimentalement les différentes molécules produites lors des étapes de recombinaison homologue.

Ces travaux sont présentés ci-après sous la forme d'un article qui a été soumis à la revue *Journal of Virology* :

P. Gayral, M. Royer and M.L. Iskra-Caruana. Evidence for activation of infectious endogenous pararetrovirus in banana (*Musa* sp.) by homologous recombination.

2.2 Article 2 : "Evidence for activation of infectious endogenous pararetrovirus in banana (*Musa* sp.) by homologous recombination"

Evidence for activation of infectious endogenous pararetrovirus in banana (*Musa* sp.) by homologous recombination

Running title: Activation of infectious BSV integrants in banana.

Philippe Gayral, Monique Royer and Marie-Line Iskra-Caruana.

CIRAD BIOS, UMR Biologie et Génétique des Interactions Plante-Parasite (BGPI)
TA A-54/K Campus international de Baillarguet, F-34398 Montpellier Cedex 5, France.

Corresponding author: M.-L. Iskra-Caruana
Tel: (+33) 4 99 62 48 13
Fax: (+33) 4 99 62 48 08
E-mail: marie-line.caruana@cirad.fr

Abstract

Banana streak virus (BSV) is a plant pararetrovirus infecting banana. BSV sequences are present as endogenous pararetroviruses (EPRVs) integrated into the nuclear genome of its host. Whereas numerous EPRVs of distinct BSV species integrated several banana host species, only a few integrants are infectious and are able to induce systemic infection by restituting a functional viral genome. This work investigates for the first time the mechanisms underlying EPRV activation. We proposed and validated a model to explain such activation based on homologous recombination (HR) of the two alleles EPRV-7 and EPRV-9 of *Banana streak goldfinger virus* (BSGFV) present in the nuclear genome of *Musa balbisiana* cv. Pisang Klutuk Wulung (PKW). Our model proposed two HR steps for EPRV-7, and three for EPRV-9, resulting in excision of a complete circular viral genome. We found evidence for these different HR events in *Escherichia coli* by analysis of the recombination of these EPRVs cloned in BAC constructs. Similar results were observed by analyzing the recombination of these BSGFV EPRVs directly in *Musa balbisiana* cv. PKW, strongly supporting that the HR may be responsible for the EPRVs BSGFV activation *in planta*. According to our model, the allele EPRV-7 is the only one generating a complete circular viral genome with any nonsense mutation, explaining previous results showing that only this allele is infectious.

Key-words: Integrated pararetrovirus, badnavirus, Banana (*Musa balbisiana*).

Introduction

The nuclear genome of numerous plants contains a large number of integrated viral sequences. These sequences belong to three viral families: *Potyviridae*, *Geminiviridae* and *Caulimoviridae*. Potyviruses have positive single stranded (ss) RNA genome ranging from 9.3 to 10.8 Kb. Endogenous potyviral sequences are found in the genome of several grapevine varieties (*Vitis vinifera*) (Tanne and Sela, 2005). Geminiviruses have small circular ssDNA genomes (2.5-3 Kb) and Geminivirus Related DNA (GRD) is found in *Nicotiana* species and Fabaceae (Murad et al., 2004; Pal et al., 2007). Caulimoviruses (i.e. plant pararetroviruses) have a circular double stranded (ds) DNA genome of about 7.5Kpb. They are found integrated into the genome of their host plant as endogenous pararetrovirus sequences (EPRVs).

EPRVs are the viral integrants the most widely distributed and the most actively studied in plants. They were found in seven different plant families: rice (*Oriza sativa*) (Kunii et al., 2004), bitter orange (*Poncirus trifoliata*) (Yang et al., 2003), dahlia (*Dahlia variabilis*) (Pahalawatta et al., 2008), petunia (*Petunia sp.*) (Richert-Pöggeler et al., 2003), potato and relatives (*Solanum sp.*) (Hansen et al., 2005; Staginnus et al., 2007b), tobacco and relatives (*Nicotiana sp.*) (Gregor et al., 2004; Jakowitsch et al., 1999; Lockhart et al., 2000) and banana (*Musa sp.*) (Geering et al., 2001; Geering et al., 2005; Ndowora et al., 1999; Safar et al., 2004).

As opposed to several animal and bacterial viruses, active and targeted integrations of plant virus DNA into the host genome have not been described so far. So far, the mechanisms underlying viral endogenization in plants remain unknown. It has been suggested that integration could originate from recombination between viral RNA genome and host mRNAs for *Potyviridae* (Maori et al., 2007; Tanne and Sela, 2005) and via illegitimate recombination targeted on micro-homologies between host and viral DNA genome for both *Geminiviridae* (Ashby et al., 1997; Bejarano et al., 1996) and *Caulimoviridae* (Staginnus and Richert-Pöggeler, 2006).

Most of the integrations occurred long ago and underwent modification since integration, resulting in both truncated viral genome and the presence of random mutations causing a loss of function of the coding viral sequences. Indeed, most of ORFs of integrated viral sequences show signs of pseudogenization with insertions, deletions of nucleotides and substitutions leading to stop codons (Geering et al., 2005; Jakowitsch et al., 1999; Kunii et al., 2004; Staginnus et al., 2007a). EPRVs are remarkable since several are able to release a functional viral genome leading to a systemic infection of the host-plant. The distinctive features of such infectious EPRVs are a non-altered coding capacity of all open reading frames (ORFs) and at least one full-length viral genome present at a single locus.

Observations of spontaneous viral infection resulting from EPRV activation were restricted to *Tobacco vein clearing virus* (TVCV) in tobacco (Lockhart et al., 2000), to *Petunia vein clearing virus* (PVCV) in petunia (Richert-Pöggeler and Shepherd, 1997), and to *Banana streak virus* (BSV) in banana (Ndowora et al., 1999).

For all these three EPRV pathosystems, interspecific hybridization is a necessary condition for EPRV activation (Lheureux et al., 2003; Lockhart et al., 2000; Richert-Pöggeler et al., 2003). For BSV-banana pathosystem, infectious EPRV are brought by the *Musa balbisiana* genome and are activated in interspecific *Musa balbisiana* x *Musa acuminata* hybrids only (Gayral et al., 2008; Lheureux et al., 2003; Ndowora et al., 1999). Three factors triggering BSV EPRV activation have been identified to date. Dahal et al., (1998) has first observed that an increase in thermal amplitude was correlated with EPRV activation. Then, the proliferation stage during micro-propagation by *in vitro* culture (i.e. the main way to multiply banana) systematically triggered EPRVs activation in interspecific *Musa acuminata* (denoted A) x *Musa balbisiana* (denoted B) hybrids (Dallot et al., 2001). Finally, Lheureux et al., (2003) observed that activation of *Banana streak Obino L'Ewai virus* (BSOLV) EPRV was dependent of a genetic factor BEL (BSV Expressed Locus) in the triploid (AAB) interspecific hybrids between virus free parents *Musa balbisiana* (BB) and *Musa acuminata* (AAAA).

The two pathosystems, PVCV in petunia and BSV in banana, are models for studies of EPRV activation in plants. PVCV EPRVs are present in 100 - 200 copies in several *Petunia* species and their potential hybrids (Richert-Pöggeler et al., 2003). The authors showed using fluorescent *in situ* hybridization that five integration loci are present in the *P. hybrida* genome. Sequence data confirmed they were made of full-length PVCV genomes arranged in tandem arrays. This head-to-tail structure would provide a DNA matrix for a direct transcription of PVCV genome from the promoter of the upstream copy to the termination site (polyadenylation site) of the downstream copy. However, such EPRVs in tandem arrays were never described for BSV.

The first description of the structure of an infectious BSV EPRV concerned the 5'-part of the integration of BSOLV in the genome of the plantain cv. Obino l'Ewai (AAB) (Harper et al., 1999). This BSV EPRV has a complex structure consisting of noncontiguous back-to-back viral sequences, interrupted by *Musa* sequences. Despite a complex organization of this integration, the full BSOLV genome was present at least once. To explain the activation from this part of EPRV, Ndowora et al. (1999) predicted a first event of homologous recombination (HR) from flanking direct repeats, to produce a terminally redundant linear BSV genome. At this step, BSOLV EPRV resembles tandem repeats of PVCV EPRVs. Then, two different mechanisms are proposed by the authors. The first one, similar to the model of activation of PVCV EPRV, involved a transcription of the integrated sequence to produce a RNA molecule that serves directly as both a viral mRNA and a template for production of viral genomic DNA by reverse-transcription. The second mechanism involves HR events between direct repeats and releases a circular BSV genome.

Despite the presence of such diversity of endogenous pararetroviruses in plants, very little is known about the mechanisms of viral integration and EPRVs activation. Infectious BSV EPRVs are a major agricultural problem. They are indeed restricted to the *M. balbisiana* genome, which is also the most promising genetic resource for breeding programs. From a theoretical point of view, they also represent a novel mode of transmission of phytoviruses. To investigate the mechanisms underlying

EPRV activation resulting in production of an infectious viral genome, we used the *Banana streak goldfinger virus* (BSGFV) integration in *Musa balbisiana* cv. Pisang Klutuk Wulung (PKW). This pathosystem was chosen for the following reasons: (i) only one copy of this BSV species exists in the genome of cv. PKW, (ii) in this cultivar, one of the two alleles (EPRV-7) described in a previous work is infectious while the other (EPRV-9) is not (Gayral et al., 2008), and (iii) the interspecific genetic cross involving cv. PKW as female parent is an adapted biological system to induce and to follow EPRV activation *in planta* in the F1 interspecific (AAB) progeny.

HR is a good candidate mechanism to explain EPRV activation. Firstly, studies of the mechanisms of recombination in plants led to the conclusion that HR is more frequent than previously suggested (Gaut et al., 2007; Hanin and Paszkowski, 2003; Li et al., 2007). Secondly, mutations resulting from HR would explain activation of EPRV frequently composed of repeated sequences. Recombination is indeed initiated by double-strand breaks that trigger intra-strand or inter-chromatid exchange between direct or inverted repeats of plant genomes (Schuermann et al., 2005). Depending on the orientations of the repeats, resolution of HR can generate several types of mutations as illustrated in Figure 1, such as insertion and deletions, inversions and excision of circular molecules.

In this study, we proposed a model to explain activation by HR of the EPRVs of BSGFV present in the nuclear genome of *Musa balbisiana* cv. PKW. In this model two HR events for EPRV-7 and three for EPRV-9 are expected to produce of a complete circular BSGFV genome. We found evidence of HR events predicted by our model in a bacterial system carrying these EPRVs cloned on a BAC as well as in *Musa balbisiana* cv. PKW.

Material and Methods

Sequence analysis

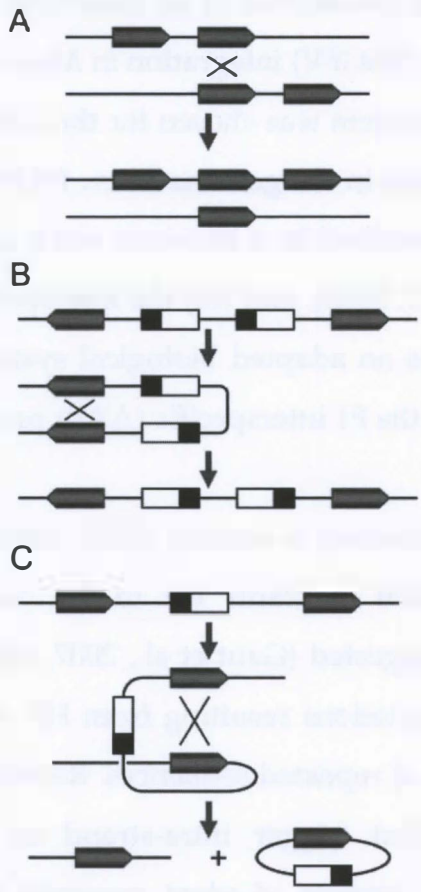

Figure 1: Schematic representations mutations caused by homologous recombinations between repeated sequences. Repeats of the same or similar sequences are shown by filled arrows. Each line represents a DNA strand from a single chromatid. **A.** Unequal crossing over between misaligned repeats on homologous chromosome or sister chromatids yields an increase in repeat number for one chromosome and a decrease in repeat number for the other chromosome. **B.** Intra-strand crossing over between inverted repeats yields an inverted fragment. **C.** Intra-strand crossing over between direct repeats causes an excision of the segment of DNA between the repeats.

BAC clones MBP_71C19 and MBP_94I16 (GenBank accession numbers AP009325 and AP009326 respectively) derived from a *Musa balbisiana* cv. PKW BAC library (Safar et al., 2004). They respectively contained the two alleles EPRV-7 and EPRV-9 annotated and characterized previously (Gayral et al., 2008). For this reason these BAC clones were called BAC7 and BAC9, respectively. EPRV sequences were edited manually using Vector NTI 10.1.1 (Invitrogen, Carlsbad, CA). The complete intergenic region of BSGFV genome (905 bp at position 6,842-483 of GenBank accession AY3509) was screened for plant cis-acting elements in the PLACE database (Higo et al., 1999) according to published data and methods (Lheureux et al., 2007; Remans et al., 2005).

DNA extraction

Total DNA was extracted by the method described in Gawel and Jarret, 1991 (Gawel and Jarret, 1991) from fresh leaf tissue of *Musa* sp. cultivars. A multiplex-immuno-capture-PCR (Le Provost et al., 2006) was used to verify the absence of BSGFV particles in *Musa balbisiana* cv. PKW and *M. acuminata* cv. Grande Naine. The quality and amount of DNA was visually estimated after separation of 5 µl of DNA extraction in a 0.8 % agarose gel, staining with ethidium bromide and visualizing under UV light.

BAC clones DNA was obtained with Wizard^R SV plus plasmid DNA purification system (Promega, Madison, WI, USA) according to the manufacturer's instructions and DNA concentration was measured with a biophotometer (Eppendorf, Hamburg, Germany).

PCR reactions, cloning and sequencing

PCRs were carried out with 5-20 ng of DNA (if not specified in the text), 20 mM Tris-HCl (pH 8.4), 50 mM KCl, 0.1 mM each dNTP, 1.5 mM $MgCl_2$, 400 nM of forward and reverse primers and 1 U Taq DNA polymerase (Eurogentec, Seraing, Belgium) in a final volume of 25µl. PCRs were performed as follow: 94 °C for 4 min (1 cycle), followed by 35 cycles at 94 °C for 30 s, 57-60 °C for 30 s, and 72 °C for 1 min/Kb and then one cycle of elongation at 72 °C for 10 min. Amplification was visualized after migration of 15 µl of PCR products on a 1.5% agarose gel in 0.5X TBE (45mM Tris-borate, 1mM EDTA, pH=8) and stained with ethidium bromide. Sequence of primers used in this study is shown in Table 1 and their location in EPRV is in Figure 2. PCR products VV2R/VM2R were cloned into TOPO-TA (Invitrogen, Carlsbad, CA) according to the manufacturer's instructions. Plasmid DNA was extracted with Wizard^R SV 96 plasmid DNA purification system (Promega, Madison, WI, USA) according to the manufacturer's instructions and sequenced with the universal primers T3 and T7. Other PCR products (DifGfF/VV4F, Gfepi-F/Gf-R and GfC/GfB) were sequenced directly in both orientations using the same primers than for PCR. Sequencing was performed by Cogenics Genome Express SA (Grenoble, France).

Results

The approach used to study EPRV activation by HR was (i) building the most parsimonious model, (ii) testing the model by checking the functionality of the excised viral genome *in silico* and (iii) testing the model experimentally by searching the products and intermediary molecules predicted by the model. In previous work, we showed that BSGFV integrant was present in a single copy with two distinct alleles (EPRV-7 and EPRV-9) in cv. PKW genome (Gayral et al., 2008). EPRV-7 and EPRV-9 alleles formed a 15 Kbp and a 17 Kbp locus, respectively, and were composed of a complex rearrangement of partially redundant viral fragments in both orientations. Six juxtaposed viral fragments - I to VI - composed EPRV-7, seven were present in EPRV-9 (Figure 2A and 2B). The general structure of both alleles was highly similar since they had five fragments in common. They however diverged by a 2.3 Kbp insertion-deletion and by small amount of substitution polymorphism (99.8 % nucleotide identity). Surprisingly, despite the structure of BSGFV integration in cv. PKW is highly rearranged, the equivalent of almost two BSGFV genomes is present in each allele.

Model of recombination

The following model was based on a minimum number of HR steps to reconstitute a full-length circular BSGFV genome from each linear allele EPRV-7 and EPRV-9. Under this model, for allele EPRV-9, a first recombination called Rec0 occurs between two direct tandem repeats of 2,300 bp present in fragment Vb and Vc (Figure 2A). In cv. PKW, Rec0 can result from both unequal reciprocal HR (Figure 1A) or from intra-molecular HR (Figure 1C). In AAB interspecific hybrids however, Rec0 can only occur from intra-molecular HR since they are hemizygote for the integration. In these hybrids, EPRV is indeed only present in *M. balbisiana* haploid set of chromosomes and absent from the two other *M. acuminata* set of chromosomes. The resolution of this recombination decreases the repeat number and releases a linear sequence with a structure similar to EPRV-7 (Figure 2B). The next steps are therefore common to EPRV-7 and EPRV-9.

Table 1: Detection of the recombination steps of EPRVs: expected and observed sizes of the PCR products amplified from the different recombined matrixes

Primers pair	Primers names	Primers sequences (5'-3')	Ta (°C)	Expected PCR product sizes (bp)			
				Native EPRVs (both alleles)	EPRV after Rec1	EPRV after Rec2 (empty locus)	EPRV after Rec2 (circular)
1	DifGfF VV4F	TTGCAGGAGCAGGAATTACA GAGCAACACGAGTCAACGAA	60	2,200	1,050[bcd]	1,050[bcd]	NA[a]
2	GfepiF Gf-R	ACGTTTCAACCCCATCAAAG TCGGTGGAATAGTCCTGAGTCTTC	57	NA[a]	700[bcd]	NA[a]	700
3	VV2R VM2R	GACAGTTCCAGCACAGCAGA TTTTGATGCATCTCCAGCAG	60	2,450	2,450	1,393[bc]	NA[a]
4	GfC GfB	CATATTCGCATTGGAAAGCA TGAGGGGACGGTCTTCTATG	60	NA[a]	NA[a]	NA[a]	1,500[bcd]

[a] No amplification
[b] PCR product observed using EPRV-7 DNA
[c] PCR product observed using EPRV-9 DNA
[d] PCR product observed using *M. balbisiana* cv. PKW total DNA

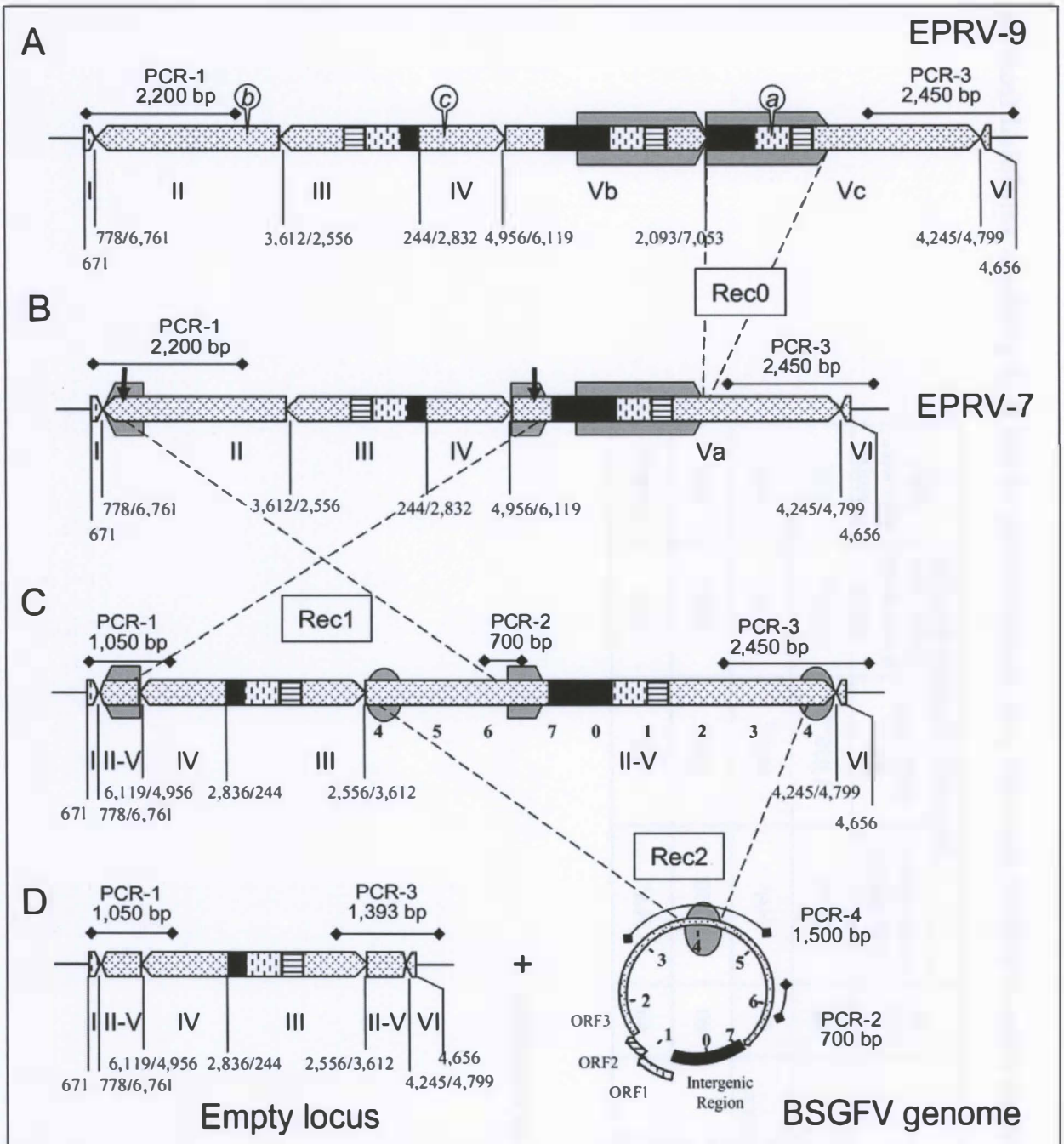

Figure 2: Model of activation of BSGFV EPRV. Borders of internal fragments of EPRV (I to V) are shown with thin vertical lines and numbers above vertical lines (Bp) and above fragments (numbered 1-7; Kbp) correspond to the position in the BSGFV genome [GenBank: AY493509]. Horizontal lines represent genomic *Musa* sequences surrounding EPRV. Repeats used as template for each HR event are indicated by the same grey symbol. Position of the 4 primer pairs used to detect HR is reported above EPRVs. **A.** Schematic representation of the structure of native BSGFV 'allele 9' found in *Musa balbisiana* cv. PKW. Letters *a, b* and *c* above EPRV represent null mutations in viral sequence. **B.** Structure of the native BSGFV 'allele 7' integrated in cv. PKW. **C.** Theoretical structure of the recombined molecule after the first event of recombination of EPRV-7 Rec1. **D.** Theoretical structure of the recombined molecule after the second recombination Rec2. Left panel: empty locus remaining in the *Musa* genome, right panel: circular molecule excised after Rec2 event and similar to the BSGFV genome.

The following HR, Rec1, occurs between two 642-bp inverted repeats separated by 6,942 bp (Figure 2B). This intra-molecular recombination results in the inversion of the internal sequence (Figure 1B) and leads to a juxtaposition of fragments II and Va for EPRV-7, and II and Vb/c for EPRV-9 as shown in Figure 2C. This intra-molecular reorganization of the EPRV generates a full-length linear BSGFV genome containing the three ORFs in the same orientation.

This viral genome is flanked by two 633-bp direct repeats, involved in the last HR step (Figure 1C). HR Rec2 produces an EPRV deleted of 7,324 bp called 'empty locus', and - more interestingly - results in the excision ('pop-out') of a circular viral molecule resembling the BSGFV circular genome (Figure 3D).

According to this model, the activation of BSGFV EPRV required two steps of HR for allele EPRV-7 and three for EPRV-9. Furthermore, only two fragments - II and V - contributed *in fine* to EPRV activation. Then, we tested the model by the analysis of EPRV sequences and experimentally.

Analysis of the open reading frames of fragments II and V

The first test of the proposed model consisted in verifying in silico the wholeness of the fragments involved in the restitution of the BSGFV genome according to our model. To do so, we analyzed the sequence of fragments II and V in EPRV-7 and EPRV-9, to verify if they were both free of any mutations causing a loss of viral function.

We checked for the presence of nonsense mutations in the coding regions of BSGFV EPRVs. For EPRV-7, all fragments were free of such mutations. In EPRV-9 however, three nonsense mutations were observed (mentioned in Figure 2A by lower case characters). Mutation *a* was an adenosine insertion A[752] in ORF1 of fragment Vc leading to a premature stop codon. Mutations *b* and *c* were substitutions creating a stop in frame (TGA) in fragments II and IV (positions A[2,632]T and C[1,933]T respectively). Consequently, the fragment II of EPRV-9, involved in the restitution of the BSGFV genome according to our model, is not functional.

The non-coding intergenic region (IR) present in fragments Va (EPRV-7), Vb and Vc (EPRV-9) contained each both the polyadenylation signal of the BSGFV circular genome (sequence AATAAT in position 114-119 on BSGFV) and the transcription

initiation site - TATA box (sequence TATATAA in position 7,100-7,106 on BSGFV) that ends the terminally redundant transcription of the full length genome typical to plant pararetroviruses. However, IR of fragment Vc of EPRV-9 was truncated and lacked the 211 first 5' nucleotides (23.3% of the total IR length), consequently, the fragment Vc of EPRV-9 involved in the restitution of the BSGFV genome according to our model is probably not functional.

Experimental validation of the model

The second test of the proposed model consisted in detecting the successive HR steps of BSGFV EPRVs according to the hypothetical model. We designed 4 PCR markers (using primers pairs 1 to 4) specific to different recombinant matrixes resulting from HR steps Rec1 and Rec2 (Table 1). These primers surrounded homologous regions involved in HR. Their relative positions on native EPRVs and on recombined matrixes are summarized in Figure 2. PCR markers were tested (i) on BAC7 and BAC9 DNA harboring EPRV-7 and EPRV-9, respectively, and (ii) on total genomic DNA of *M. balbisiana* cv. PKW. PCR products were sequenced in both orientations and electrophoregrams were checked for ambiguous base calls using Staden Package software (Staden et al., 1999). Sequencing was performed directly on PCR products, however cloning was required for PCR using primer pair 3 (see below) due to low amplification yield.

PCR using primer pair 1 detects the 5' repeat region of the recombined EPRV matrix resulting from Rec1 recombination, by amplifying a 1,050 bp product. This PCR also amplifies a 2,200 bp product on native EPRV-7 and EPRV-9. This 1,050 bp product was successfully detected in Figure 3A (line 3 and 4) showing that recombination was detectable when a large amount of BAC DNA were used as PCR template (> 10^4 BAC copies), whatever the EPRV concerned (EPRV-7 or EPRV-9). At lower DNA concentrations (Figure 3A lines 1 and 2) this PCR product was almost undetectable. Although this PCR product was not visible *in planta* (cv. PKW) on this gel (Figure 3A line 5), other independent experiments have proved the opposite (data not shown). Cloning and sequencing of this PCR product amplified either from BAC clones and

from cv. PKW confirmed our prediction: the sequence started with the fragment I and ended with the fragment IV on opposite orientation, instead of fragment II.

PCR using primer pair 2 was designed to amplify a single 700 bp product specific to the 3' repeat region of the recombined EPRV matrix resulting from Rec1 recombination. Once again, the expected PCR product was observed for both BAC clones, and cv. PKW (Figure 3B lines 3-4 and 5, respectively). As previously described, cloning and sequencing of these products confirmed our model.

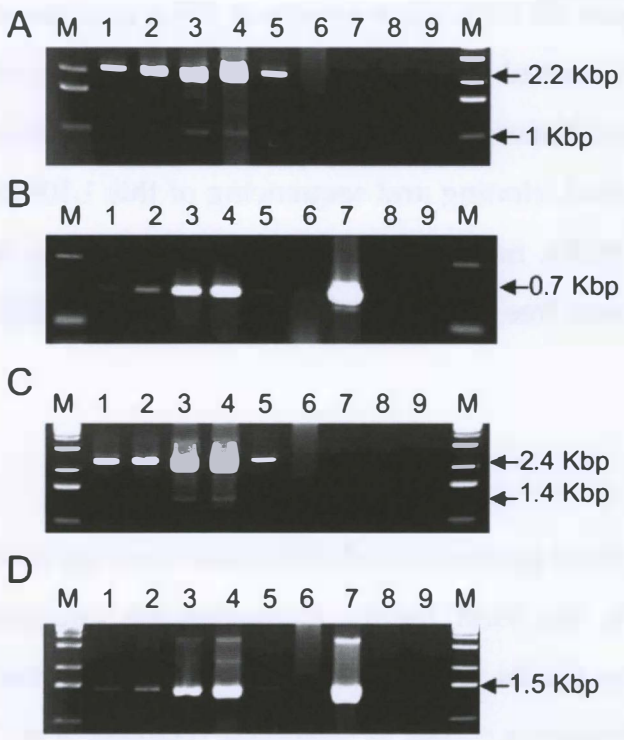

Figure 3: Detection of recombined BSGFV EPRV in BAC clones and *in planta*. A: PCR using primer pair 1: detection of the recombined EPRV matrix resulting from Rec1 recombination (left repeat region, 1,050 bp product), **B:** PCR using primer pair 2: detection of Rec1 recombination of EPRV (right repeat region, 700 bp product), **C:** PCR using primer pair 3: detection of the second HR step Rec2 (3' part of the empty locus, 1,393 bp product), **D:** PCR using primer pair 4: detection of the BSGFV genome released after the HR step Rec2 (circular form, 1,500 bp product). Lane M: 1Kb ladder (Invitrogen, Carlsbad, CA), Lanes 1 and 3: BAC7. Lanes 2 and 4: BAC9. Lanes 1 and 2: PCR with 6 ng DNA (approx. 10 BAC copies) and lanes 3 and 4: 80 ng DNA (approx. 10^4 BAC copies). Line 5 and 6: total genomic DNA of *M. balbisiana* cv. PKW with 5 ng and 100 ng DNA, respectively. Line 7: total genomic DNA of *M. acuminata* cv. Grande Naine infected by BSGFV virus, L8 total DNA of *M. acuminata* cv. Grande Naine BSGFV-free and L9 water control.

PCR using primer pair 3 was designed to amplify a single 1,393 bp product resulting from the second HR step Rec2. The PCR product was observed in both BAC clones but not in cv. PKW (Figure 3C, lines 3 and 4). As previously described, cloning and sequencing confirmed the 3' part of the empty locus represented in Figure 3D.

The last PCR using primer pair 4 was specific to the circular form of the BSGFV genome and could detect the circular viral genome released after the HR step Rec2 according to our model of activation. According to this model, this PCR amplifies a 1,500 bp product located on ORF3 of the virus. As expected, this 1,500 bp product was observed in Figure 3D with *Musa acuminata* DNA infected with episomal BSGFV virus as a positive control (lane 7). Once again, the recombined product was observed in both BAC clones (lines 3 and 4) and also *in planta* (i.e. cv. PKW) (line 5). As previously described, cloning and sequencing of this 1,500 bp product confirmed our model. For all PCRs, no amplification was observed in the negative controls: DNA of *Musa acuminata* free of episomal and endogenous BSGFV (line 8); and water (line 9).

Discussion

Recent advances in plant genomics and molecular virology revealed the presence of sequences related to the viral family *Caulimoviridae* integrated into the nuclear genome of numerous plants. Several EPRVs colonized efficiently the host genome and reach several thousands copies in *Solanaceae* (Hansen et al., 2005; Lockhart et al., 2000; Matzke et al., 2004; Richert-Pöggeler et al., 2003; Staginnus et al., 2007a). Although the vast majority of integrated viral sequences lost their potential harmful effect for the host plant through pseudogenization, some integrants however kept the ability to induce infection. To date, only three pathosystems are known to harbor infectious EPRVs but further research may extend the list. Viral propagation through activation of infectious BSV is not only a major issue in banana culture and breeding, but is also an original replication strategy of plant viruses never described before. For the first time, we proposed and validated a model based on HR to explain the activation of *Banana streak Goldfinger virus* EPRV in *Musa balbisiana* cv. PKW.

We constructed the activation model using the previously characterized BSGFV EPRV locus in *Musa balbisiana* cv. PKW composed of two alleles EPRV-7 and EPRV-9 (Gayral et al., 2008). In this previous work, we showed that two fragments - II and V- covered completely the BSGFV genome [see Figure 4 in Gayral et al., (2008)]. In the present study, these two fragments appeared to be sufficient to reconstruct *in silico* a circular molecule resembling the BSGFV genome in a minimum of HR steps. The use of any other fragments or supplementary steps of HR would undeniably increase the probability of adding deleterious mutations, and decrease the probability of activation. We also verified that the repeat regions involved in HR are long enough to enable recombination. It has been shown that a minimum of 352 bp repeat is sufficient to induce spontaneous HR in tobacco (Puchta, 2000; Zubko et al., 2000). The length of the two repeat regions in our model (> 600 bp) is thus theoretically adequate for recombination. Finally, we confirmed that for the allele EPRV-7, the regions involved in the restitution of BSGFV did not contain any deleterious mutations for the viral function of the restituted BSGFV genome.

During multi-template PCR or when several annealing sites are present in the template, such as for EPRVs, a chimera could be formed from an incompletely extended primer (Shuldiner et al., 1989) and template switching (Odelberg et al., 1995). However, three lines of evidence diminished the PCR artifact possibility for HR events detection by PCR in BAC7, BAC9 and in *M. balbisiana* cv. PKW. First, it has been reported that the rate of artificial recombination occurring among related DNA sequences during PCR was low: 5 % (Meyerhans et al., 1990); a more recent study estimated this rate between 2.5 to 8.7 % depending on the Taq-polymerase used (Qiu et al., 2001). Second, since template switching would be a random event, the breakpoints positions would not have necessarily followed the positions predicted by our model. Third, template switching and re-annealing in homologous region within EPRV would result in a population of different chimeras displaying length polymorphism, which had not been detected by direct sequencing of PCR products.

In 1991, Gal and co-workers used a stable transgene integration of *Cauliflower mosaic virus* (CaMV) in *Brassica napus* to study the mechanisms of homologous recombination (Gal et al., 1991). The transgene was a linear CaMV genome with terminal direct repeats of 1,033 bp long thus strongly resembling the BSGFV EPRV-7 following the first step of recombination Rec1. Spontaneous HR between the repeat regions occurred in a limited number of cells and led to the excision of a circular CaMV genome, resulting in a systemic CaMV infection in the plants. This artificial CaMV EPRV provides further evidence that activation of EPRV in plants via HR likely exists, provided that repeat regions surround a full-length viral genome.

The bacterial system used in this study has proved to be a useful model to detect the four recombined regions predicted from our hypothetical model using PCR of EPRV-7 and EPRV-9. The first two primer pairs (1 and 2) aimed to detect the first HR event of EPRV-7 (Rec1) by amplifying the recombined molecules present in all subsequent recombined molecules. This HR step (Rec1) produces the contiguous BSGFV genome flanked with direct repeats. The last two primers aimed to detect the output of the proposed model: the empty locus resulting from the excision ('pop-out') of the contiguous BSGFV genome, as well as the excised circular fragment representing an infectious BSGFV genome. This latter viral form would then produce pregenomic transcript, subsequently translated to produce viral proteins and reverse transcribed to replicate the viral DNA genome. In contrast, using *M. balbisiana* genome as DNA template, we failed to detect the 3' part of the empty locus resulting from HR step Rec2. Since large amount of template DNA was required to have a sufficient signal in our PCR experiments, this absence of detection lack would probably result from the limited sensitivity of the method regarding too few DNA templates. The use of larger amount of total DNA of *M. balbisiana* cv. PKW in PCR was not successful because it contributed to inhibit the reaction. Furthermore, reproducibility of detecting Rec1 using primer pair 1 from *M. balbisiana* DNA was low. Altogether, these results suggest that the amount of recombined matrix in our experiments remained extremely low, in line with expected rare recombination events.

Previous attempt to explain EPRV activation proposed different molecular mechanisms according to the structure of the integration locus. For instance, the structure of endogenous PVCV in the petunia genome (ePVCV) is a tandem array of PVCV genomes. Its activation is suspected to result solely from the direct transcription of ePVCV, resulting in the production of a pre-genomic RNA that would initiate the viral infection (Richert-Pöggeler et al., 2003). The structures described for BSV EPRV are however much more complex, and a direct transcription would not release a functional pre-genomic RNA. HR was thus suggested to provide the required EPRV modifications to release a functional viral genome (Ndowora et al., 1999), and our results support this hypothesis. However, the transcription of EPRV following a HR event may also explain EPRV activation (Ndowora et al., 1999). Further studies are required to assess whether both HR and transcription mechanisms also occur in BSGFV EPRV and in other infectious BSV EPRVs species present in the cv. PKW genome such as BSGFV EPRV and BSImV EPRV (Iskra-Caruana, unpublished).

We strongly suspected that EPRV-9 is not functional. In a previous study, EPRV-7 was correlated with the release of BSGFV in the F1 interspecific triploid (genotype AAB) progeny between *Musa balbisiana* (BB) and *Musa acuminata* cv. IDN1104x (AAAA), while EPRV-9 was never associated with infectious BSGFV (Gayral et al., 2008). The present study provides further evidence indicating that BSGFV genome resulting from EPRV-9 activation would contain corrupted ORFs whereas those from EPRV-7 would be functional. Despite EPRV-9 is not infectious, we did not observe any difference regarding production of recombined molecules from the two alleles. These results suggest that, even if HR events are scarce, they are not a limiting factor for EPRV activation. On the opposite, the coding capacity of the ORFs (i.e. the absence of deleterious mutations) seems determinant for distinguishing between infectious and non-infectious EPRVs. We are confident that only very rare HR events along the single infectious EPRV copy present in the genome of *M. balbisiana* cv. PKW are sufficient to produce a limited number of BSGFV genome. The restituted BSGFV genomes are infectious, since they induce a systemic infection in the interspecific triploid (AAB) progeny of cv. PKW carrying EPRV-7. In cv. PKW

however, we showed that EPRV activation by recombination seems to occur, and that a BSGFV genome was likely restituted. The most surprising is that cv. PKW remains BSV-free following activator stresses such as *in vitro* culture. It is also resistant to BSGFV infection after inoculation of BSGFV particles by its insect vector (Mealybug, *Planoccocus citri*) (Iskra Caruana M.L. et al., May 28-31, 2003). The resistance toward activation of endogenous BSGFV sequences in cv. PKW is therefore more likely due to a repression of BSGFV replication rather than EPRV recombination. Future studies that aimed to investigate BSV resistance in *Musa balbisiana* should therefore focus on the factors controlling virus multiplication rather than solely on the first molecular steps of EPRV activation. Recent advanced on epigenetic silencing mechanism of other plant-EPRV pathosystems might enlighten this issue (Staginnus and Richert-Pöggeler, 2006). Transcriptional gene silencing (TGS) by DNA methylation, chromatin modification and siRNAs production from EPRV would prevent EPRV activation by maintaining low transcription levels of ePVCV and LycEPRV (EPRV sequences closely related to TVCV in the tomato genome) (Noreen et al., 2007; Staginnus et al., 2007a). Additionally, siRNAs would also prevent the multiplication of any episomal virus by post-transcriptional gene silencing (PTGS), provided they share enough homologies with EPRV sequences (Mette et al., 2002). This epigenetic model explaining both the control of EPRV activation by plants and an EPRV-mediated virus resistance in *Solanaceae* has yet to be validated in banana.

Acknowledgements

We are very grateful to Serge Galzi, Laurence Blondin and Nathalie Laboureau for technical assistance. We thank Matthieu Chabannes and Gaël Thébaud for their helpful comments. Materials and methods fulfilled the requirements of quality management system according to the guidelines of ISO 9001:2000 standard. P. G. was supported by a 'Cirad - Région Languedoc Roussillon' PhD grant.

Literature cited

Ashby, M.K., Warry, A., Bejarano, E.R., Khashoggi, A., Burrell, M. and Lichtenstein, C.P. (1997) Analysis of multiple copies of geminiviral DNA in the genome of four closely related *Nicotiana* species suggest a unique integration event. Plant. Mol. Biol. 35(3), 313-21.

Bejarano, E.R., Khashoggi, A., Witty, M. and Lichtenstein, C. (1996) Integration of multiple repeats of geminiviral DNA into the nuclear genome of tobacco during evolution. Proc. Natl. Acad. Sci. USA 93(2), 759-64.

Dahal, G., Hughes, d.A., Thottappilly, G. and Lockhart, B.E.L. (1998) Effect of temperature on symtom expression and reliability of *banana streak badnavirus* detection in naturally infected plantain and banana (*Musa* spp). Plant Dis. 82, 16-21.

Dallot, S., Acuna, P., Rivera, C., Ramirez, P., Cote, F., Lockhart, B.E.L. and Caruana, M.L. (2001) Evidence that the proliferation stage of micropropagation procedure is determinant in the expression of *Banana streak virus* integrated into the genome of the FHIA 21 hybrid (*Musa* AAAB). Arch. Virol. 146(11), 2179-2190.

Gal, S., Pisan, B., Hohn, T., Grimsley, N. and Hohn, B. (1991) Genomic homologous recombination in planta. Embo J. 10(6), 1571-8.

Gaut, B.S., Wright, S.I., Rizzon, C., Dvorak, J. and Anderson, L.K. (2007) Recombination: an underappreciated factor in the evolution of plant genomes. Nat. Rev. Genet. 8(1), 77-84.

Gawel, N.J. and Jarret, R.L. (1991) A modified CTAB DNA extraction procedure for *Musa* and *Ipomea*. Plant Mol. Biol. Rep. 9, 262-266.

Gayral, P., Noa-Carrazana, J.-C., Lescot, M., Lheureux, F., Lockhart, B.E.L., Matsumoto, T., Piffanelli, P. and Iskra-Caruana, M.-L. (2008) A single *Banana streak virus* integration event in the banana genome as the origin of infectious endogenous pararetrovirus. J. Virol. 82(13), 6697-6710.

Geering, A.D.W., Olszewski, N.E., Dahal, G., Thomas, J.E. and Lockhart, B.E.L. (2001) Analysis of the distribution and structure of integrated Banana streak virus DNA in a range of Musa cultivars. Molecular Plant Pathology 2(4), 207-213.

Geering, A.D.W., Olszewski, N.E., Harper, G., Lockhart, B.E.L., Hull, R. and Thomas, J.E. (2005) Banana contains a diverse array of endogenous badnaviruses. J. Gen. Virol. 86, 511-520.

Gregor, W., Mette, M.F., Staginnus, C., Matzke, M.A. and Matzke, A.J.M. (2004) A distinct endogenous pararetrovirus family in Nicotiana tomentosiformis, a diploid progenitor of polyploid tobacco. Plant Physiology 134(3), 1191-1199.

Hanin, M. and Paszkowski, J. (2003) Plant genome modification by homologous recombination. Curr. Opin. Plant Biol. 6(2), 157-162.

Hansen, C.N., Harper, G. and Heslop-Harrison, J.S. (2005) Characterisation of pararetrovirus-like sequences in the genome of potato (*Solanum tuberosum*). Cytogenet. Genome Res. 110(1-4), 559-565.

Harper, G., Osuji, J.O., Heslop-Harrison, J.S.P. and Hull, R. (1999) Integration of *Banana streak badnavirus* into the *Musa* genome: molecular and cytogenetic evidence. Virology 255(2), 207-213.

Higo, K., Ugawa, Y., Iwamoto, M. and Korenaga, T. (1999) Plant cis-acting regulatory DNA elements (PLACE) database: 1999. Nucleic Acids Res. 27(1), 297-300.

Iskra Caruana M.L., Lheureux F., Noa-Carrazana J.C., Piffanelli P., Carreel F., Jenny C., Laboureau N. and Lockhart B.E.L. (May 28-31, 2003) Unstable balance of relation between pararetrovirus and its host plant: the BSV-EPRV banana pathosystem, In : Abstracts : EMBO Workshop Genomic Approaches in Plant Virology, pp. 8, Keszthely, Hungary.

Jakowitsch, J., Mette, M.F., van der Winden, J., Matzke, M.A. and Matzke, A.J.M. (1999) Integrated pararetroviral sequences define a unique class of dispersed repetitive DNA in plants. Proc. Natl. Acad. Sci. USA 96(23), 13241-13246.

Kunii, M., Kanda, M., Nagano, H., Uyeda, I., Kishima, Y. and Sano, Y. (2004) Reconstruction of putative DNA virus from endogenous rice tungro bacilliform virus-like sequences in the rice genome: implications for integration and evolution. BMC Genomics 5(1), 80.

Le Provost, G., Iskra-Caruana, M.L., Acina, I. and Teycheney, P.Y. (2006) Improved detection of episomal Banana streak viruses by multiplex immunocapture PCR. J. Virol. Methods 137(1), 7-13.

Lheureux, F., Carreel, F., Jenny, C., Lockhart, B. and Iskra-Caruana, M. (2003) Identification of genetic markers linked to banana streak disease expression in inter-specific *Musa* hybrids. Theor. Appl. Genet. 106(4), 594-598.

Lheureux, F., Laboureau, N., Muller, E., Lockhart, B.E. and Iskra-Caruana, M.L. (2007) Molecular characterization of *banana streak acuminata Vietnam virus* isolated from *Musa acuminata siamea* (banana cultivar). Arch. Virol. 152(7), 1409-16.

Li, J., Hsia, A.P. and Schnable, P.S. (2007) Recent advances in plant recombination. Curr. Opin. Plant Biol. 10(2), 131-135.

Lockhart, B.E., Menke, J., Dahal, G. and Olszewski, N.E. (2000) Characterization and genomic analysis of *tobacco vein clearing virus*, a plant pararetrovirus that is transmitted vertically and related to sequences integrated in the host genome. J. Gen. Virol. 81, 1579-1585.

Maori, E., Tanne, E. and Sela, I. (2007) Reciprocal sequence exchange between non-retro viruses and hosts leading to the appearance of new host phenotypes. Virology 362(2), 342-9.

Matzke, M., Gregor, W., Mette, M.F., Aufstatz, W., Kanno, T., Jakowitsch, J. and Matzke, A.J.M. (2004) Endogenous pararetroviruses of allotetraploid Nicotiana tabacum and its diploid progenitors, N. sylvestris and N. tomentosiformis. Biological Journal of the Linnean Society 82(4), 627-638.

Mette, M.F., Kanno, T., Aufsatz, W., Jakowitsch, J., van der Winden, J., Matzke, M.A. and Matzke, A.J.M. (2002) Endogenous viral sequences and their potential contribution to heritable virus resistance in plants. Embo Journal 21(3), 461-469.

Meyerhans, A., Vartanian, J.P. and Wain-Hobson, S. (1990) DNA recombination during PCR. Nucleic Acids Res. 18(7), 1687-91.

Murad, L., Bielawski, J.P., Matyasek, R., Kovarik, A., Nichols, R.A., Leitch, A.R. and Lichtenstein, C.P. (2004) The origin and evolution of geminivirus-related DNA sequences in *Nicotiana*. Heredity 92(4), 352-8.

Ndowora, T., Dahal, G., LaFleur, D., Harper, G., Hull, R., Olszewski, N.E. and Lockhart, B. (1999) Evidence that badnavirus infection in *Musa* can originate from integrated pararetroviral sequences. Virology 255(2), 214-220.

Noreen, F., Akbergenov, R., Hohn, T. and Richert-Pöggeler, K.R. (2007) Distinct expression of endogenous *Petunia vein clearing virus* and the DNA transposon dTph1 in two *Petunia hybrida* lines is correlated with differences in histone modification and siRNA production. Plant J. 50(2), 219-229.

Odelberg, S.J., Weiss, R.B., Hata, A. and White, R. (1995) Template-switching during DNA synthesis by *Thermus aquaticus* DNA polymerase I. Nucleic Acids Res. 23(11), 2049-57.

Pahalawatta, V., Druffel, K. and Pappu, H. (2008) A new and distinct species in the genus *Caulimovirus* exists as an endogenous plant pararetroviral sequence in its host, *Dahlia variabilis*. Virology 376(2), 253-257.

Pal, A., Chakrabarti, A. and Basak, J. (2007) New motifs within the NB-ARC domain of R proteins: Probable mechanisms of integration of geminiviral signatures within the host species of *Fabaceae* family and implications in conferring disease resistance. J. Theor. Biol. 246(3), 564-573.

Puchta, H. (2000) Removing selectable marker genes: Taking the shortcut. Trends Plant Sci. 5(7), 273-274.

Qiu, X., Wu, L., Huang, H., McDonel, P.E., Palumbo, A.V., Tiedje, J.M. and Zhou, J. (2001) Evaluation of PCR-Generated Chimeras, Mutations, and Heteroduplexes with 16S rRNA Gene-Based Cloning, pp. 880-887. Vol. 67.

Remans, T., Grof, C.P.L., Ebert, P.R. and Schenk, P.M. (2005) Identification of functional sequences in the pregenomic RNA promoter of the *Banana streak virus* Cavendish strain (BSV-Cav). Virus Res. 108(1-2), 177-186.

Richert-Pöggeler, K.R., Noreen, F., Schwarzacher, T., Harper, G. and Hohn, T. (2003) Induction of infectious petunia vein clearing (pararetro) virus from endogenous provirus in petunia. Embo J. 22(18), 4836-4845.

Richert-Pöggeler, K.R. and Shepherd, R.J. (1997) *Petunia vein-clearing virus*: A plant pararetrovirus with the core sequences for an integrase function. Virology 236(1), 137-146.

Safar, J., Noa-Carrazana, J.C., Vrana, J., Bartos, J., Alkhimova, O., Sabau, X., Simkova, H., Lheureux, F., Caruana, M.L., Dolezel, J. and Piffanelli, P. (2004) Creation of a BAC resource to study the structure and evolution of the banana (Musa balbisiana) genome. Genome 47(6), 1182-1191.

Schuermann, D., Molinier, J., Fritsch, O. and Hohn, B. (2005) The dual nature of homologous recombination in plants. Trends Genet. 21(3), 172-181.

Shuldiner, A.R., Nirula, A. and Roth, J. (1989) Hybrid DNA artifact from PCR of closely related target sequences. Nucleic Acids Res. 17(11), 4409.

Staden, R., Beal, K.F. and Bonfield, J.K. (1999) The Staden Package, 1998, Bioinformatics Methods and Protocols, pp. 115-130.

Staginnus, C., Gregor, W., Mette, M.F., Teo, C.H., Borroto-Fernandez, E.G., Machado, M.L., Matzke, M. and Schwarzacher, T. (2007a) Endogenous pararetroviral sequences in tomato (*Solanum lycopersicum*) and related species. BMC Plant Biol. 7, 24.

Staginnus, C., Gregor, W., Mette, M.F., Teo, C.H., Borroto-Fernandez, E.G., Machado, M.L., Matzke, M. and Schwarzacher, T. (2007b) Endogenous pararetroviral

sequences in tomato (Solanum lycopersicum) and related species. BMC Plant Biol 7, 24.

Staginnus, C. and Richert-Pöggeler, K.R. (2006) Endogenous pararetroviruses: two-faced travelers in the plant genome. Trends Plant Sci. 11(10), 485-491.

Tanne, E. and Sela, I. (2005) Occurrence of a DNA sequence of a non-retro RNA virus in a host plant genome and its expression: evidence for recombination between viral and host RNAs. Virology 332(2), 614-22.

Yang, Z.N., Ye, X.R., Molina, J., Roose, M.L. and Mirkov, T.E. (2003) Sequence analysis of a 282-kilobase region surrounding the citrus Tristeza virus resistance gene (Ctv) locus in Poncirus trifoliata L. Raf. Plant Physiol 131(2), 482-92.

Zubko, E., Scutt, C. and Meyer, P. (2000) Intrachromosomal recombination between attP regions as a tool to remove selectable marker genes from tobacco transgenes. Nat. Biotechnol. 18(4), 442-445.

CHAPITRE II

Diversité et évolution des EPRV

et de la diversification des hôtes *Musa*. Cette étude nous a permis d'estimer le nombre d'évènements d'intégrations indépendants qui se sont produits dans le génome des bananiers.

Poue finir, nous avons analysé les vitesses d'évolution et le type d'évolution moléculaire de ces séquences afin de comparer les pressions de sélection agissant sur les séquences virales libres et sur les EPRV.

Ces travaux sont présentés ci-après sous la forme d'un article soumis à la revue *Molecular Phylogenetics and Evolution* (actuellement en reviewing) :

P. Gayral and M.L. Iskra-Caruana. Phylogeny of Banana streak virus reveals a recent burst of integrations in the genome of banana (*Musa* sp.).

1.2 Article 3: "Phylogeny of *Banana streak virus* reveals a recent burst of integrations in the genome of banana (*Musa* sp.)"

Phylogeny of *Banana streak virus* reveals a recent burst of integrations in the genome of banana (*Musa* sp.)

Philippe Gayral and Marie-line Iskra-Caruana.

CIRAD BIOS, UMR Biologie et Génétique des Interactions Plante-Parasite (BGPI) TA A-54/K Campus international de Baillarguet, F-34398 Montpellier Cedex 5, France.

Corresponding author:

M.-L. Iskra-Caruana

Tel: (+33) 4 99 62 48 13

Fax : (+33) 4 99 62 48 08

E-mail: marie-line.caruana@cirad.fr

Abstract

Banana streak virus (BSV) is a plant dsDNA pararetrovirus (family *Caulimoviridae*, genus *badnavirus*). Although integration is not an essential step in the replication cycle of BSV, the nuclear genome of bananas and plantains (genus *Musa*) contains BSV endogenous pararetrovirus sequences (BSV EPRVs). Some of these are infectious by reconstituting a functional viral genome. Recent studies revealed a large molecular diversity of episomal BSV viruses (i.e. non-integrated) while other studies focused on BSV EPRV sequences (i.e. integrated). The evolutionary history of *badnavirus* integration in *Musa* sp. was inferred from phylogenetic relationships between BSV and BSV EPRVs. We also compared the relative evolution rates and selective pressures (d_N/d_S ratio) between integrated and episomal viral sequences. At least 27 recent independent integration events occurred after the divergence of three *Musa* species, indicating that viral integration is a recent and frequent phenomenon. A strong relaxation of selective pressure on badnaviral sequences that experienced neutral evolution after integration in the plant genome was recorded. Additionally, we observed a significant 35% decrease in EPRV evolution rate compared to BSV, reflecting the difference in evolution rate between episomal dsDNA viruses and plant genome. The comparison of our results with the evolution rate of *Musa* genome and other reverse-transcribing viruses suggests that EPRVs played an active role in episomal BSV diversity and evolution.

Key-words

Banana (*Musa sp.*), *Banana streak virus* (BSV), Endogenous pararetrovirus (EPRV), Integrated badnavirus, d_N/d_S ratio, Molecular evolution rate, Selective constraints.

Introduction

Plants nuclear genomes harbor a vast number and diversity of endogenous viral sequences. Interestingly, none of the plant viruses described up to now require an obligatory integration step for replication. Endogenous pararetrovirus sequences (EPRVs) are an important type of integrated viral sequences. EPRVs are related to the family *Caulimoviridae*, also called plant pararetroviruses (PRV), which have a circular double stranded dsDNA genome (7.0 to 8 Kbp). In this family composed of 6 genera, EPRVs are described in the genus *Petuvirus*, *Cavemovirus*, *Badnavirus*, *Tungrovirus* and *Caulimovirus*. EPRVs are scattered among distantly related plant families, and were found in the nuclear genome of bitter orange (*Poncirus trifoliata*), rice (*Oriza sativa*), potato and relatives (*Solanum* sp.), petunia (*Petunia* sp.), tobacco and relatives and banana (*Musa* sp.) (reviewed in Staginnus and Richert- Pöggeler 2006), and in lucky bamboo (*Dracaena sanderiana*) (Su et al. 2007) and Dahlia (*Dahlia variabilis*) (Pahalawatta et al. 2008).

Because copy number of EPRV in tobacco reached up to thousands, they were described as a novel class of dispersed repetitive elements with significant impact on host genomes complexity and evolution (Hohn et al. 2008; Jakowitsch et al. 1999). All described EPRVs have a similar rearranged pattern with tandem repeats, internal duplications, fragmentations and inversions of viral genomes (Gayral et al. 2008; Ndowora et al. 1999; Richert- Pöggeler et al. 2003). EPRVs are hypothesized to originate from illegitimate or non-homologous recombination between plant and episomal (*i.e.* non-integrated) viral genomes during the first step of infection when viruses enter the nucleus (Staginnus and Richert- Pöggeler 2006). The majority of EPRVs result in partial and non-functional viral genomes (Geering et al., 2005a; Kunii et al., 2004). However, few integration events result in functional ORFs containing a full-length viral genome. Such uncorrupted EPRV sequences can then be activated, resulting in the release of functional viral genomes infecting the host plant. Infectious EPRVs were reported for *Petunia vein clearing virus* (PVCV) in Petunia - *Petunia hybrida* (Richert- Pöggeler and Shepherd 1997), for *Tobacco vein clearing virus* (TVCV) in tobacco - *Nicotiana tabacum* (Lockhart et al.

2000), and for *Banana streak virus* (BSV) in banana - *Musa sp.* (Ndowora et al. 1999). Two main mechanisms are proposed to explain EPRVs activations: the homologous recombination between repeat regions surrounding EPRV resulting in the excision of a circular viral genome (Gaut et al. 2007; Ndowora et al. 1999; Schuermann et al. 2005) and the transcription of EPRVs leading to a viral pregenomic RNA (Noreen et al. 2007; Richert- Pöggeler et al. 2003).

BSV are non-enveloped bacilliform viruses causing banana streak disease in all the banana-producing areas (Lockhart and Jones 2000). This pathogen appears to be a very useful model to study viral evolution and the consequences of integration causing rapid changes in selective pressures. Indeed, BSV exists as an episomal pathogen species transmitted horizontally by mealybugs and infecting wild and domesticated banana species (genus *Musa*), but also as EPRV sequences in the *Musa* genome. Studies of episomal virus sequences during epidemics in Uganda (Harper et al. 2004; Harper et al. 2005), Australia (Geering et al. 2000) and Mauritius (Jaufeerally-Fakim et al. 2006) revealed a great genetic diversity among BSV. Based on the sequence analysis of a partial (580 bp) reverse transcriptase (RT) /RNase H gene located in ORF3, and obtained by PCR using degenerate primers, the 'Caulimoviridae study group' of the International Committee on Taxonomy of Viruses (ICTV) defined a 20% nucleotide diversity in the RT/RNase H region as the threshold to distinguish between two episomal badnavirus species (Fauquet et al. 2005). BSV is therefore composed of several distinct virus species able to infect a same *Musa* host plant, rather than several strains of the same viral species. Furthermore, no significant genetic exchanges were detected between BSV species (Muller E., pers. comm.), suggesting the presence of species barrier. At the same time, Geering et al., (2005a) searched for badnavirus sequences integrated in the *Musa* genome by a degenerate PCR approach with virus-free plant material. They also observed a high diversity of endogenous BSV-related sequences present in several *Musa* species. Nevertheless, no phylogenetic studies investigating the relationships between episomal virus particles and EPRV sequences were performed so far. This step is critical to understand the evolutionary history of

badnaviruses considered as emerging pathogens in tropical countries, and the phenomenon of viral integration in plants. In this study, the terms BSV EPRVs and badnavirus EPRVs will always refer to integrated viral sequences whereas BSV, badnavirus and PRV (pararetrovirus) will refer to episomal viruses.

The first goal addressed in this study was to infer a robust phylogeny and to revise the taxonomy of the genus. For the first time, all published episomal and EPRV sequence were used in a same dataset: 13 species of badnavirus, 99 EPRV sequences from several *Musa* species and genotypes and 105 BSV sequences of uncertain origin. Furthermore, we wanted to know if badnavirus phylogeny was congruent with a specialization on a given plant species, and if a co-diversification between BSV and banana could be observed. We then used this phylogenetic framework to address evolutionary issues of EPRVs. What is the distribution of integrated sequences: are they dispersed all along the genetic diversity of BSV or do they rather cluster in a small number of BSV groups? Are integration events rare or frequent (*i.e.* how many independent integration events occurred in the *Musa* genome)? Then, are these integration events ancient (*i.e.* are they shared between *Musa* species) or rather recent (*i.e.* restricted to a single *Musa* species but shared between sub-species or genotypes)? Finally, the effects of integration events on viral DNA evolution were studied by comparing the evolution rate and selective pressure acting on episomal sequences - evolving in a host-parasite interaction context, *vs.* EPRVs - being a part of the host genome.

Materials and Methods

Nucleotide sequences

A 540 bp fragment of the RT/RNase H region located in ORF3 in the genome of badnaviruses was used in this study (Table 1). The sequences correspond either to episomal viruses and were labeled 'PRV', or to integrated sequences and were labeled 'EPRV'. The 'PRV' category is composed of sequences retrieved from full length published viral genomes of both BSV and six closely related badnaviruses species. In this study, we generated one additional sequence of the episomal

Banana streak cavendish virus (BSCavV) (GenBank accession numbers shown in Table 1). This sequence was generated from a BSCavV-infected *Musa acuminata* cv. Williams after a multiplex-immuno-capture-PCR (Le Provost et al. 2006) that amplified episomal viral particles only with primers BadnaFP 5'GCCITTYGGIITIAARAAYGCICC3' and BadnaRP 5'CCAYTTRCAIACISCICCCCAICC3' (Yang et al. 2003) at annealing temperature (Ta) of 55°C. Product was cloned and 13 positive clones were sequenced using M13F universal primer. The sequences showed more than 98% similarity and a single sequence was chosen in this study.

Sequences of the 'EPRV' category are endogenous sequences from both BSV and badnaviruses. 87 EPRV sequences downloaded from GenBank were generated from a degenerate PCR approach using total DNA of virus-free *Musa acuminata*, *Musa balbisiana* and *Musa schizocarpa* accessions (Geering et al. 2005a). Eleven additional EPRV sequences were generated in this study from *Musa acuminata* and *Musa balbisiana* checked for BSV-free status by multiplex-immuno-capture-PCR (Le Provost et al. 2006). PCR amplification with total DNA of *Musa balbisiana* cv. Pisang Klutuk Wulung (PKW) was performed using primers Badna1 5'CTNTAYGARTGGYTNGTNATGCCNTTYGG3' and GfR 5'TCGGTGGAATAGTCCTGAGTCTTC3' at Ta = 51 °C and with 25 pmol (instead of 10 pmol) of primer Badna1 in the PCR mix, and the product was cloned. Two clones (PKW514 and PKW515) were sequenced in both orientations with the universal primers M13F and M13R (GenBank accession numbers shown in Table 1). PCR amplification with DNA from *Musa acuminata* cv. Grande Naine (GN) was performed using primers BadnaFP and BadnaRP described above, and the product were cloned. Nine clones (FP2, FP4, FP6, FP14, FP19, FP20, FP22, FP26 and FP28) were sequenced in both orientations with primers M13F and M13R (GenBank accession numbers shown in Table 1). An additional BSGFV EPRV sequence was retrieved from the bacterial artificial chromosome (BAC) clone MBP_71C19 (GenBank AP009325) made with the genome of *Musa balbisiana* cv. PKW (Gayral et al. 2008).

A third category (labeled 'BSUgV') contains sequences with an undetermined status (PRV or EPRV). This dataset of 105 sequences was obtained from Harper et al., (2005). The authors screened mainly *Musa acuminata* plants from BSV epidemics in Uganda and used a degenerate immunocapture (IC)-PCR approach (see discussion). All available BSUgV sequences were downloaded from GenBank and were named according to information provided by GenBank rather than by the original publication: BSUgV species E in the publication refers to species E and F in GenBank, and BSUgV species F to L in the publication refers to species G to M in GenBank.

In this study, total plant and viral genomic DNA was extracted from leaf tissue following a previously described method (Gawel 1991). PCRs were carried out with 5-20 ng of DNA, 20 mM Tris-HCl (pH 8.4), 50 mM KCl, 0.1 mM of each dNTP, 1.5 mM $MgCl_2$, 10 pmol of each primer and 1 U Taq DNA polymerase (Eurogentec) in 25 µl. PCRs were performed by first heating at 94 °C for 4 min, followed by 35 cycles at 94 °C for 30 s, 51-55 °C for 30 s, and 72 °C for 1 min/Kb and then one cycle of elongation at 72 °C for 10 min. PCR products were cloned into pGEM-T easy (Promega) or TOPO-TA (Invitrogen) vectors according to manufacturer's instructions. Plasmid DNA was extracted with Wizard^R SV plus plasmid DNA purification system (Promega) according to the manufacturer's instructions. Sequencing was performed by Cogenics Genome Express SA (Grenoble, France).

Table 1: Categories of sequences used in this study

Category	Sequence name	Clone name and GenBank accession number	Reference
BSUgV	BSUgBV	BSUgBV115 (AJ968463)	Harper et al., 2005
	BSUgCV	BSUgCV114 (AJ968464)	
	BSUgDV	BSUgDV521 (AJ968465)	
	BSUgEV	BSUgEV112 and -523 (AJ968466, AJ968467)	
	BSUGFV	BSUGFV113 (AJ968469)	
	BSUgGV	BSUgGV532 (AJ968471)	
	BSUgHV	BSUgHV221 (AJ968472)	
	BSUgIV	BSUgIV14, -15, -16 (AJ968475 to AJ968477), BSUgIV172, -173, -191 (AJ968492 to AJ968494), BSUgIV193, -210, -212, -31, -36 (AJ968495, AJ968481, AJ968483 to AJ968485), BSUgIV421, -422, 423, -432, -436 (AJ968496 to AJ968500) BSUgIV45, -51, -53, -56, AJ968488, AJ968489 to AJ968491)	
	BSUgJV	BSUgJV293 and -296 (AJ968502, AJ968503)	
	BSUgKV	BSUgKV81, -82, -94 (AJ968504, AJ968505, AJ968507)	
	BSUgLV	BSUgLV222, -271, -283, -285, -32, -333, -344, -346 (AJ968517 to AJ968520, AJ968508, AJ968523 to AJ968525), BSUgLV610, -611, -73, -74, -76, -83, -84 (AJ968510 to AJ968516)	
	BSUgMV	BSUgMV102 , -104, -132, -135, -143, -162, (AJ968526, AJ968527, AJ968529 to AJ968531, AJ968533), BSUgMV164, -165, -215, -263, -301, -302, (AJ968534, AJ968535, AJ968539 to AJ968542), BSUgMV321, -364, -365, -372, -373, -381, -382 (AJ968544, AJ968547, AJ968548, AJ968550, AJ968551), BSUgMV381, -382, -383, -511, -515, -516 (AJ968553 to AJ968558)	
	BSUgAV	BSUgAV445, 452, -456, -466, -472, -473, -481, -482 (AJ968454, AJ968455, AJ968457 to AJ968462)	
	BSUgGFV	BSUgGFV55, -542, -544, -545, -546, -548, (AJ968437 to AJ968442)	
	BSUgImV	BSUgImV11, -26, -232, -391, -492, -496 (AJ968444, AJ968446, AJ968448 to AJ968451)	
	BSUgOLV	BSUgOLV154, -171, -181, -182, -231, -244, -284, -311, -322, (AJ968422 to AJ968430), BSUgOLV342, -42, -43 (AJ968432 to AJ968434), BSUgOLV43 and -44 (AJ968419, AJ968420)	
Category	Banana host species	Clone name and GenBank accession number	Reference
EPRV	*Musa schizocarpa* (SS)	Shiz2, Shiz3, Shiz25, Shiz14, Shiz23, Shiz24 (Accession number AY189378 to AY189383)	Geering et al., 2005
	Musa acuminata subsp. *banksii* (AA)	Bank1, Bank10, Bank11, Bank13, Bank14, Bank17, Bank19, Bank6, Bank8 (AY189384 to AY189392), Bank9 (AY452278)	

	Musa acuminata subsp. *burmannicoides* (AA) cv. 'Calcutta 4'	Cal12, Cal13, Cal1, Cal22, Cal27, Cal30, Cal34, Cal6, Cal8, Cal22t (AY189444 to AY189453)	
	Musa acuminata subsp. *malaccensis* (AA)	Mal10, Mal11, Mal15, Mal22, Mal26, Mal3, Mal6, Mal8, (AY189393 to AY189400)	
	Musa balbisiana cv. 'Pisang Batu' (BB)	Bat10, Bat19, Bat20, Bat21, Bat24, Bat25, Bat27, Bat2, Bat31, Bat34, Bat36, Bat4, Bat5, Bat6, Bat8, Bat9 (AY189420 to AY189435)	
	Musa balbisiana cv. PKW (BB)	PKW12, PKW16, PKW18, PKW23, PKW32, PKW36, PKW8, PKW9 (AY189436 to AY189443)	
	cv. 'Obino l'Ewai' (genotype AAB)	OBLE15, OBLE17, OBLE1, OBLE21, OBLE24, OBLE2, OBLE32, OBLE34, OBLE35, OBLE36, OBLE37, OBLE3, OBLE4, OBLE5, OBLE7, OBLE8, OBLE13t, OBLE1t, OBLE22t (AY189401 to AY189419)	
	cv. 'Klue Tiparot' (genotype ABB)	KT11, KT23 (AY452259 and AY452260), KT30, KT31, KT32, KT36, KT37, KT38, KT42, KT51, KT6, KT9 (AY452262 to AY452271)	
	Musa acuminata cv. 'grande Naine' (AAA)	FP2, FP4, FP6, FP14, FP19, FP20, FP22, FP26, FP28 (EU908850 to EU908858)	This study
	Musa balbisiana cv. PKW (BB)	PKW514, PKW515 (EU908849, available upon acceptance of the manuscript)	This study
	Musa balbisiana cv. PKW (BB)	EPRVGF	Gayral et al., 2008

Category	Species name	Accession number
PRV	*Taro bacilliform virus* (TabV)	AF357836
	Citrus yellow mosaic virus (CMBV)	AF347695
	Cacao swollen shoot virus (CSSV)	L14546
	Kalanchoe top-spotting virus (KTSV)	AY180137
	Commelina yellow mottle virus (ComYMV)	X52938
	Sugarcane bacilliform Mor virus (SCBMV)	M89923
	Banana streak Obino l'Ewai virus (BSOLV)	NC_003381
	Banana streak Mysore virus (BSMyV)	NC_006955
	Banana streak Acuminata Vietnam virus (BSAcVnV)	AY750155
	Banana streak Acuminata Yunnan virus (BSAcYuV)	DQ092436
	Banana streak Imove virus (BSImV)	unpublished data
	Banana streak Goldfinger virus (BSGFV)	AY493509
	Banana streak Cavendish virus (BSCavV)	This study (EU908859)

Phylogenetic inference

Sequences were aligned using ClustalW (Thompson et al. 1994) implemented in Bioedit (Hall 1999) and corrected manually when necessary.

The software DAMBE version 4.5.20 (Xia and Xie 2001) was used to detect substitution saturation in each of the six alignments following a previously described method (Xia et al. 2003). For this purpose, the percentage of invariant sites was first estimated by DAMBE with a Poisson + Inv. distribution with the defaults parameters. For this analysis, the sequence of the outgroup *Taro bacilliform virus* (TaBV) was excluded from the alignments. Expected saturation index was given for asymmetric tree topology and estimated for 16 OTU after 500 replicates.

Before inferring phylogenies, we used Modeltest 3.7 (Posada and Crandall 1998) to choose the evolutionary model that best fitted our data using Akaike informative criterion (AIC), since AIC has been shown to have advantages in model selection over hierarchical likelihood ratio tests (Posada and Buckley 2004). For each alignment, the best model and associated parameters were then used to infer tree topologies by maximum likelihood using PAUP 4.0b10 (Swofford 2002). Branching supports were assessed by performing 500 bootstrap replicates using PHYML (Guindon and Gascuel 2003).

At the same time, a bayesian approach was also used to confirm the topologies of each alignment previously inferred with the ML approach, using the software MrBayes 3.1.2 (Huelsenbeck and Ronquist 2001). Each run was performed with 5 chains and 10^6 generations using the default priors of the GTR model. Bayesian posterior probabilities were calculated from majority-rule consensus of trees sampled every 20 generations once the Markov chains had become stationary (determined by empirical checking of likelihood values).

Estimation of evolution rate and selective pressures

The phylogenetic trees were labeled to differentiate the branches leading to 'PRV', to 'EPRV' or to '*BSUgV*' sequences. Since episomal viral sequences can integrate and conversely EPRV can be activated, the status of a taxa or lineage can change through

time *i.e.* along the branches of the phylogenetic tree. The real status of a sequence is therefore known only at the time of its sampling. To minimize the biases caused by a wrong classification of sequences, only the terminal branches where labeled. Furthermore, our sampling showed a sufficient diversity to provide many short terminal branches.

Variations in synonymous and non-synonymous substitution rates for each sequence category were analyzed using the *branch models* of program codeml implemented in PAML 3.15 (Yang 1997; Yang 1998). To compare the d_N/d_S ratio between 'PRV' and 'EPRV' for instance, two models were build. A null model (M0) assumes a common d_N/d_S ratio for 'PRV' and 'EPRV' terminal branches, plus an second ratio for the remaining branches *i.e.* the internal branches and the 'BSUgV' branches'. M0 was compared to an alternative model assuming one ratio for 'PRV' and one for 'EPRV', and one for the remaining branches. The likelihood ratio of the two models to be compared (M0 versus M1) tested whether the alternative model fits the data significantly better than the null hypothesis: twice the difference in log likelihood between the two models is compared with a X^2 distribution with n degrees of freedom, n being the difference between the numbers of parameters of the two models. Stop codons were found in several EPRV sequences and were removed prior to PAML analysis. Inserted nucleotides were removed; deleted nucleotides and the third base of substituted stop codons were coded as unknown. The estimations of variations of d_N/d_S ratio between integrated and non-integrated viral sequences were used as a proxy for investigating the selective forces occurring before and after BSV integrations.

In order to verify that an increase of non-synonymous/synonymous (d_N/d_S) ratio was only due to a relaxation of selective constraints, rather than to a positive selection effect, the presence of positively selected amino acids was tested by the *site models* of program codeml implemented in PAML (Yang 1998; Yang and Nielsen 1998). The neutral model (M7) uses a discrete beta distribution [range (0, 1)] to model different d_N/d_S ratios between sites. The alternative model M8 assumes a supplementary class of codons with $d_N/d_S > 1$. M7 and M8 were compared with likelihood ratio tests (LRT).

The program baseml of PAML implements local clock models which assume that branches of the tree can be partitioned into several rate groups (Yang and Yoder 2003; Yoder and Yang 2000). As for d_N/d_S ratio comparison, the last branches of the tree were again labeled in categories to test if PRV and EPRV had a same or a distinct substitution rate relatively to the rate observed in the rest of the tree. Neutral models assuming a first rate for 'PRV' and 'EPRV' branches plus a rate for the other branches (M0: two-ratio models) were compared by LRTs with alternative models assuming a rate for the branches of the category 'EPRV', a rate for 'PRV' and a rate of the other branches (M1: three-ratio models). A change in molecular evolution rate was examined between BSV and BSV EPRVs in order to propose an evolutionary scheme of the integration event.

Results

Analysis of the phylogenetic signal

To test if our alignments were suitable for phylogenetic studies, we tested the presence of substitution saturation with the program DAMBE. It is assumed that phylogenetic information is essentially lost when the observed saturation index is equal or larger than half of full substitution saturation (Xia 1999). Software DAMBE estimated the expected saturation indices assuming half of the full substitution saturation and compared it to the observed saturation indices. Substitution saturation is detected when the observed indices are higher than the expected indices. We tested the presence of saturation in each of the six alignments (see below) used in this study (Alignments in supplementary data). No group showed signs of saturation ($p < 0.03$), therefore validating our data set for phylogenetic analyses (Table 2).

Overall phylogeny of BSVs

The overall phylogeny was inferred from 25 sequences chosen as a representative sample of the diversity of BSV sequences and close badnavirus species (named *overall PRV* alignment) and is shown in Figure 1. Three deeply rooted groups were distinguished and named groups 1, 2 and 3. These groups were supported by higher

bootstrap values in phylogenies with additional sequences (see Figure 2-4). However, the order of emergence of the three groups is still not clear as suggested by the basal trifurcation.

All previously described species of BSV clustered in group 1: BSV species Obino l'Ewai (BSOLV) (Harper and Hull 1998), BSV species Golfinger (BSGFV) (Gayral et al. 2008), BSV species Mysore (BSMyV) (Geering et al. 2005b), BSV species Imové (BSImV) (Gayral et al., unpublished data) and BSV species acuminata Vietnam (BSAcVNV) (Lheureux et al. 2007).

Table 2: Detection of saturation substitution with the program DAMBE

Alignment [a]	Observed Saturation index (Iss)	Expected Saturation Index (Iss.cAsym)	T	DF	*p*-value [b]
Overall PRV	0.458	0.518	2.16	473	0.0317
Group 1	0.445	0.518	2.58	434	0.0102
Group 2	0.355	0.516	6.24	506	0.0000
Group 3	0.373	0.517	5.54	427	0.0000
Group 1 + 2	0.448	0.519	2.15	500	0.0319
Main BSUgV	0.435	0.517	3.11	427	0.0020

[a] Alignments detailed in the results section

[b] The statistical significance of the difference between observed and expected saturation indexes was assessed with a two-tailed test

Three sub-groups were also observed in this group (Figure 2), the first one is represented by BSMyV species, the second subgroup contained BSOLV and BSV species Cavendish (BSCavV) species, the newly discovered sequence BSUgAV and the kalanchoe top-spotting badnavirus (KTSV) isolated on *Kalanchoe blossfeldiana*. The last subgroup comprised banana-infecting species only: BSGFV, BSV species Yunnan (BSAcYuV), BSAcVNV, and BSImV.

Groups 2 and 3 encompassed closely related badnavirus species, but mainly BSUgV sequences collected from infected banana cultivars during Uganda epidemics (Harper et al. 2005). Group 2 contained seven taxa of BSUgV (from B to H) and *Citrus mosaic bacilliform virus* (CMBV) and *Cacao swollen shoot virus* (CSSV) infecting citrus and cocoa (*Theobroma cacao*), respectively, were in basal position. Group 3 contained both the other BSUgV taxa (from I to M) and *Sugarcane bacilliform mor virus* (SCBMV)

infecting sugarcane (*Saccharum officinarum*). *Commelina yellow mottle virus* (ComYMV) infecting *Commelina sp* was in basal position in this third group.

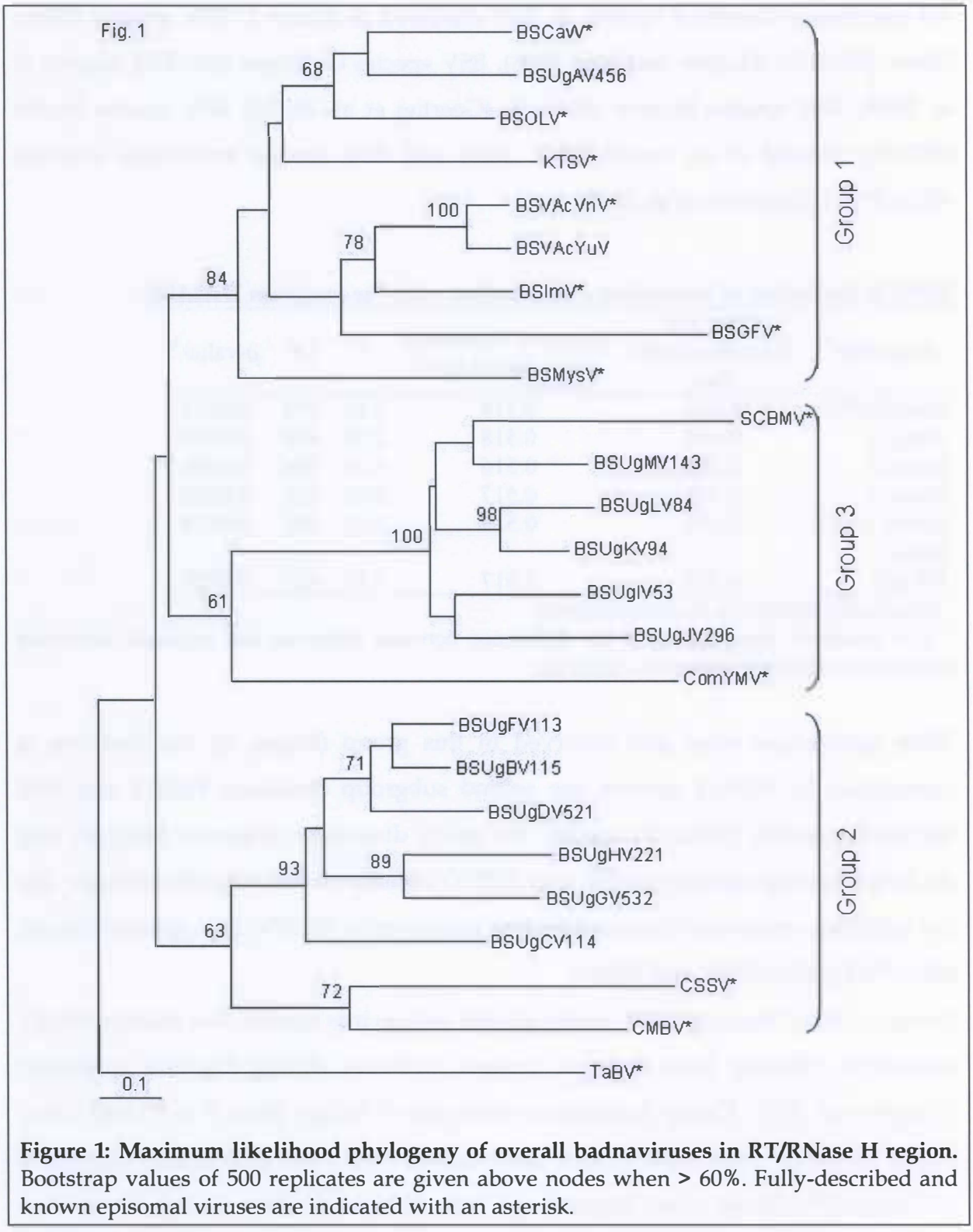

Figure 1: Maximum likelihood phylogeny of overall badnaviruses in RT/RNase H region. Bootstrap values of 500 replicates are given above nodes when > 60%. Fully-described and known episomal viruses are indicated with an asterisk.

Origin and evolution of BSV integrations

A ML phylogeny based on the initial dataset containing all 217 sequences was first reconstructed and the same tree topology as in Figure 1 was obtained (data not shown). To facilitate further analyses, this initial tree was split in three parts, according to the three major phylogenetic groups. *Alignment 1, alignment 2* and *alignment 3* contained 76, 86 and 72 sequences, respectively, and encompassed all the sequences that belong to groups 1, 2 and 3, respectively, as well as sequences from the other two groups as outgroups. All sequences were labeled as either 'PRV' for episomal viruses, 'EPRV' for integrated sequences or 'BSUgV' for sequences of Uganda origin with unclear status (integrated or episomal). To study the distribution of integration events in the *Musa* genome, it was necessary to distinguish between two independents events. An independent event was defined when one or more EPRV sequences forming a single and well supported phylogenetic group were separated from other EPRVs by episomal sequences.

Group 1 (Figure 2) encompassed EPRV, PRV, as well as BSUgV sequences. Except BSUgV species A (denoted BSUgAV) that formed a new phylogenetic group not closely related to any known BSV species, all other BSUgV sequences corresponded to already described BSV species such as BSOLV, BSImV and BSGFV. We recorded at least ten independent integration events in this group; the majority (6/10) were restricted to the B genome and noted EPRV-B in Figure 2. They occurred in diploid *M. balbisiana* genotypes (BB) such as cv. PKW and cv. Pisang batu, but also in interspecific genotypes such as cv. Klue tiparot (ABB) and cv. Obino l'Ewai (AAB). At least two integrations were found in A x B interspecific hybrids. As these EPRV sequences did not cluster near EPRV-A or EPRV-B groups, their attribution to the A or B genome remained undetermined. Finally, two integrations were specific to the A genome (denoted EPRV-A) and appeared in *Musa acuminata* (AA) only. It is not clear whether the following sequences integrated in the A genome (clones Cal1, Cal22 and Cal8) and in the genome A or B (clones OBLE5 and KT30) derived from the same integration event or from two independent events, since they are not separated by episomal sequences.

Fig. 2

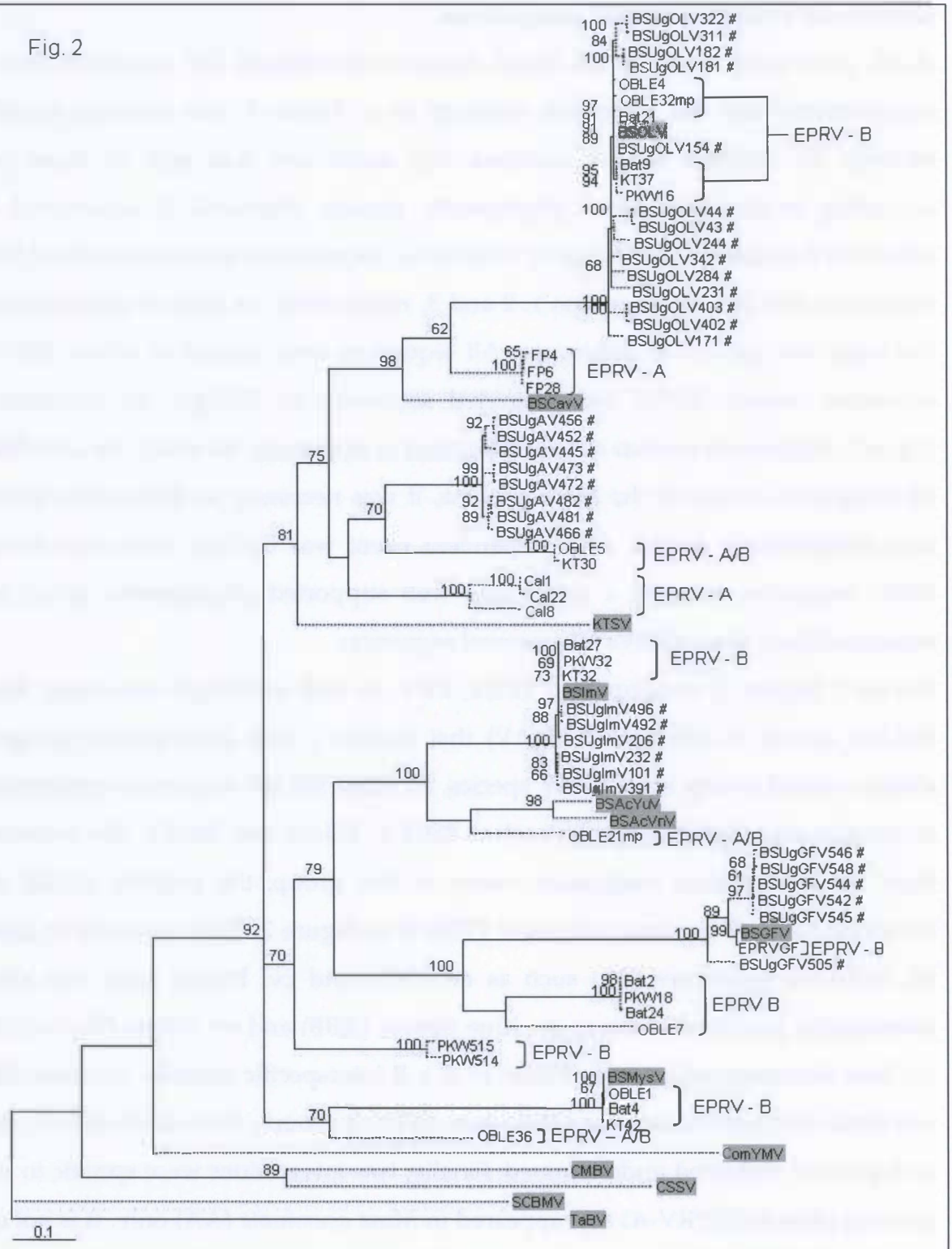

Figure 2: Maximum likelihood phylogeny of episomal and integrated BSV of group 1. The phylogeny is based on a 540 bp alignment of RT/RNase H viral region. Bootstrap values of 500 replicates are given when > 60%. Other sequences of group 2 and group 3 are given as outgroups. BSUgV sequences are indicated by a hash sign. Episomal sequences of both BSV and badnaviruses are shaded. EPRV-A and EPRV-B are sequences integrated in the *Musa acuminata* genome (denoted A) and the *M. balbisiana* genome (denoted B), respectively. EPRV-A/B sequences are found in interspecific AxB *Musa* genotypes and can not be assigned to a particular genome when not close to a given phylogenetic group.

Nevertheless, the deep divergence observed between the two groups (Cal8-KT30: 0.61 substitution/site) suggests that two independent integrations occurred. Figure 2 illustrates that EPRVs related to BSOLV, BSImV, BSGFV and BSMyV are associated to the B genome while EPRVs related to BSCavV, BSAcYuV, BSAcVNV are relatively more associated to the A genome. EPRVs corresponding to unknown viruses are observed in both A and B genomes. Furthermore, we did not observe any EPRV corresponding to known and unknown viruses integrated into the *Musa schizocarpa* genome.

Group 2 contained the largest part of EPRV analyzed in this study, as well as seven of the newly discovered BSUgV species (namely species B to H) infecting bananas. Seven subgroups were defined (named 2a to 2g) (Figure 3). Two subgroups diverged early during the evolution of group 2. The first one, subgroup 2g, contained two EPRVs (PKW9 and KT36) found in the B genome. The second corresponded to two badnaviruses that do not infect *Musa*: CMBV and CSSV (subgroup 2f). It is not clear whether subgroup 2g diverged first, since even if the node was well supported by bootstrap value in the 'Alignment 2' phylogeny (77%), it was much less supported (bootstrap value = 34%) in the bayesian approach (branch trifurcation; data not shown), as well as in the ML phylogeny in the larger dataset 'Alignment 1+2' (data not shown). This latter alignment is a merge of *alignment 1* and *alignment 2* in a single dataset of 154 sequences and encompassed the only two groups containing EPRV sequences; group 3 was indeed free of EPRV (see below).

The count uncertainty for group 2 is likely to be higher than for group 1 since fewer reliable episomal virus clustered in this group. However, several BSUgV sequences that separate EPRV sequences were probably episomal viruses. Furthermore, the great divergence between groups of EPRV sequences often enables to discriminate between integration events (see below). Sequences of group 2 corresponded to integrated sequences within the three banana species studied: *Musa balbisiana*, *M. acuminata* and *M. schizocarpa* (genome S).

Fig. 3

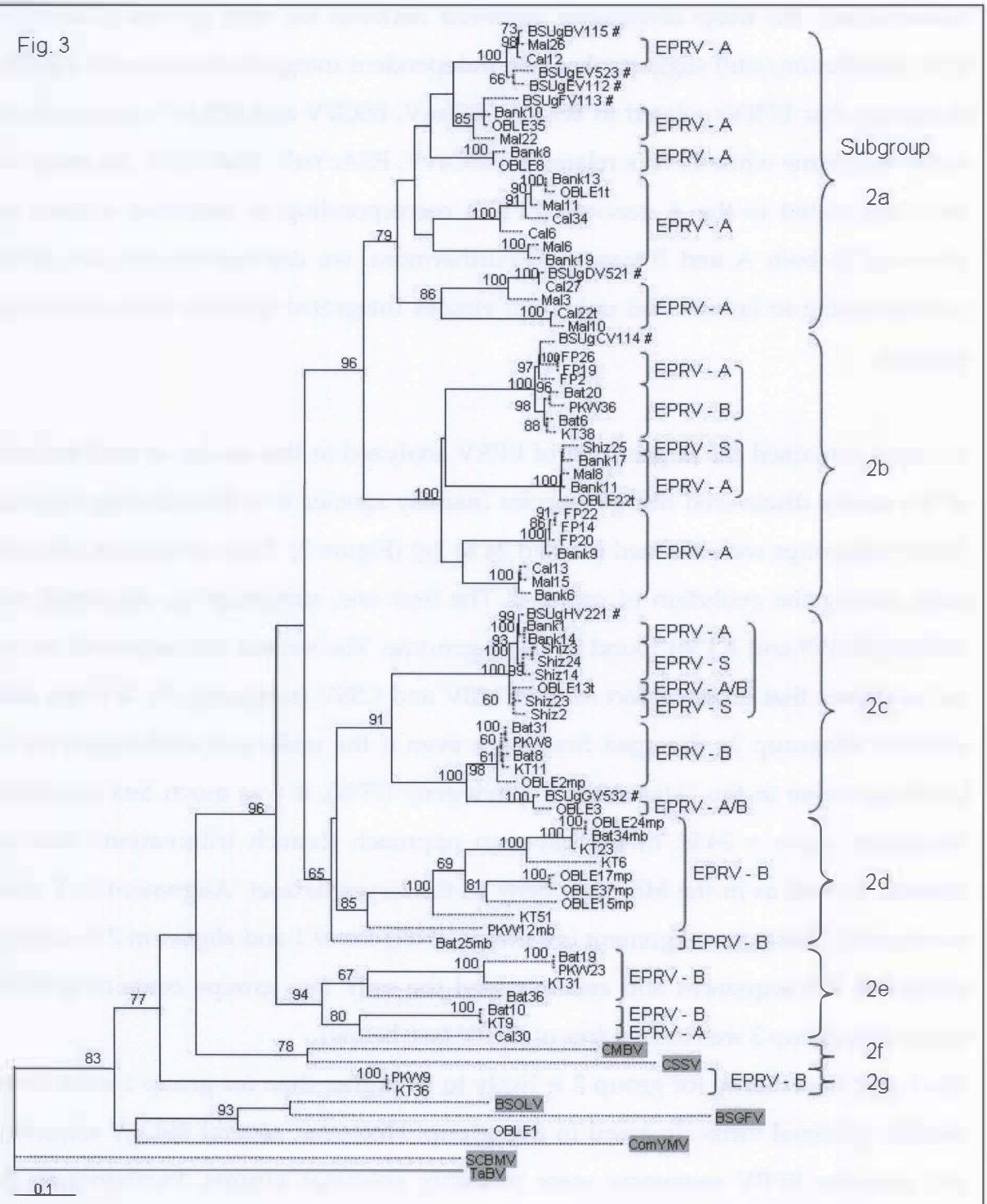

Figure 3: Maximum likelihood phylogeny of episomal and integrated PRV of group 2. The phylogeny is established with a 540 bp alignment of the RT/RNase H viral region. Bootstrap values of 500 replicates are given when > 60%. Other sequences of groups 1 and 3 are given as outgroups. BSUgV taxa are indicated by a hash sign. Episomal BSV and episomal badnaviruses are shaded. EPRV-A, EPRV-B and EPRV-S are sequences integrated in the *Musa acuminata* (A), *M. balbisiana* (B) and *M. schizocarpa* (S) genomes, respectively. EPRV-A/B sequences are found in interspecific AxB *Musa* genotypes and can not be assigned to a particular genome when not close to a given phylogenetic group.

In several subgroups, EPRVs sequences belonged to a single *Musa* species: subgroup 2a integrated the A genome only, subgroup 2d and 2g integrated the B genome only, suggesting that they originated from independent integration events. All the sequences deriving from a single integration event are therefore paralogues. The situation was different for subgroups 2b, 2c and 2e since sister groups were described in different *Musa* species. In subgroup 2e, one group integrated the B genome and its sister group the A genome, but the two sister groups showed a great divergence (0.4 substitution/site). In subgroups 2b and 2c however, the topology was similar: the two sister groups integrated in genome A and B (subgroup 2b) and in genome A and S (subgroups 2b and 2c), but in this case the divergence between the sister groups was very small (0.08, 0.04 and 0.03 substitution/site respectively). In three cases, the integration events occurred most likely before the speciation of the *Musa* genus so they are relatively ancient, each sister group containing orthologous sequences. Altogether, at least 17 independent integration events were found in group 2: seven were specific to A genome, six to B genome, one was undetermined (either in A or in B) one was common to A and B and two were common to *Musa* genomes A and S.

Figure 4 shows the ML phylogeny of group 3. Surprisingly, this group was free of any described EPRV and was mainly composed of BSUgV sequences. A first subgroup diverged first and is represented by the sole SCBMV sequence, a virus infecting sugarcane. The three other subgroups are composed of BSUgV sequences (species I to M). For all alignments, bayesian reconstructions produced congruent topology with the ML inferences, at the level of general topology as well as for terminal branching of trees (data not shown).

Rate of evolution and selection of BSV and EPRV

Status of BSUgV sequences from Uganda epidemics

The phylogeny of integrated and non-integrated sequences revealed that EPRV sequences did not form a limited phylogenetic group, but were dispersed among two clades (group 1 and group 2). It is therefore not possible to assign the category of a given sequence solely by its phylogenetic position. We addressed this question for BSUgV sequences distributed among the three groups and whose category was not

clearly known. Their molecular evolution pattern was studied in order to determine if they evolved as PRV or EPRV sequences. Table 3 shows for each phylogenetic group the estimations and comparison by LRT of the d_N/d_S ratio and evolution rate of the terminal branches of different sequences categories: 'PRV', 'EPRV' and 'BSUgV'.

For alignments corresponding to groups 1, 2, 3 and combination of 1 and 2, BSUgV terminal branches had a d_N/d_S ratio significantly different from the one of the PRV terminal branches. The d_N/d_S ratio was 30 times higher for BSUgV than PRV in group 2, and 4 times higher in group 1 and 3. In order to test if this increase in the d_N/d_S ratio could be the result of positive selection acting on some codons rather than a relaxation of selective constraints, signals of positive selection were searched in the BSUgV sequences using PAML. We defined a new alignment containing 41 of the 105 published BSUgV sequences representative of the genetic diversity of BSUgV observed in the three phylogenetic groups, as alignment *Main BSUgV*. No positive selection was found with this new dataset (Table 3: M7 vs. M8, p=1). Conversely, the d_N/d_S ratio of BSUgV branches was not significantly different from those of EPRV branches in all phylogenetic groups. BSUgV sequences contained a molecular evolution signal closer to EPRV than to PRV, and this result was mostly observed in group 2. It is therefore possible that integrated sequences exist among BSUgV sequences previously defined as PRV sequences. BSUgV sequences will therefore remain in our datasets since they obviously contribute to BSV phylogeny, but they will be categorized as neither EPRV nor PRV in the next steps.

Evolution of EPRV and PRV terminal branches

The evolutionary pattern between terminal branches of EPRV and those of PRV were compared using LRT. Table 4 shows the estimates of d_N/d_S ratio and relative rates of evolution of terminal branches of PRVs and EPRVs, for each phylogenetic group where integration occurred (groups 1 and 2). In all phylogenetic groups, the evolution speed of EPRV terminal branches was significantly lower than those of PRVs, with a factor rEPRV/rPRV of about 0.6 for group 1 and group 2.

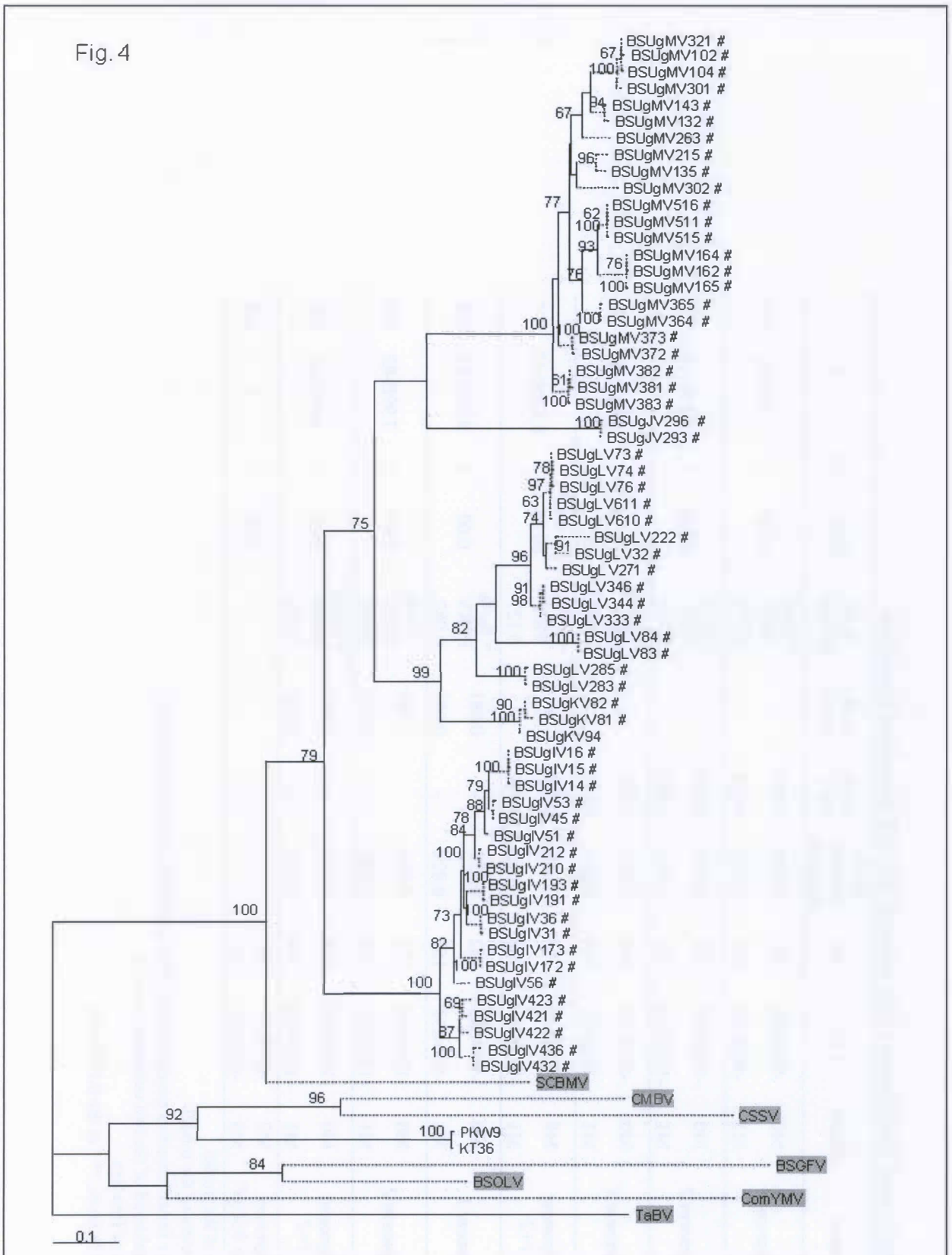

Figure 4: Maximum likelihood phylogeny of episomal and integrated PRV of group 3. The phylogeny is established with a 540 bp alignment of the RT/RNase H viral region. Bootstrap values of 500 replicates are given when > 60%. Other sequences of groups 1 and 2 are given as outgroups. BSUgV taxa are indicated by a hash sign. Episomal sequences of both BSV and badnaviruses are shaded.

Table 3: d_N/d_S ratio and likelihood ratio tests of BSUgV terminal branches

Comparison	Dataset	Model	LnL	np	d_N/d_S internal branches	d_N/d_S PRV	d_N/d_S EPRV	d_N/d_S BSUgV	2dLnL	df	*p*	
BSUgV vs. PRV	*Alignment 1*	M0	-9620.48	152	0.040	0.042	-	= d_N/d_S PRV	23.63	1	1.13E-06	***
		M1	-9608.66	153	0.040	0.018	-	0.079				
	Alignment 2	M0	-13608.87	172	0.061	0.052	-	= d_N/d_S PRV	62.42	1	2.78E-15	***
		M1	-13577.66	173	0.061	0.009	-	0.283				
	Alignment 3	M0	-8029.23	144	0.035	0.052	-	= d_N/d_S PRV	10.19	1	1.41E-03	**
		M1	-8024.13	145	0.034	0.014	-	0.063				
	Alignment 1+2	M0	-19285.55	308	0.056	0.054	-	= d_N/d_S PRV	60.25	1	8.330E-15	***
		M1	-19255.42	309	0.055	0.016	-	0.127				
BSUgV vs. EPRV	*Alignment 1*	M0	-9604.32	152	0.027	-	0.081	= d_N/d_S EPRV	0.00	1	9.61E-01	NS
		M1	-9604.32	153	0.027	-	0.081	0.082				
	Alignment 2	M0	-13518.25	172	0.029	-	0.188	= d_N/d_S EPRV	3.07	1	7.99E-02	NS
		M1	-13516.71	173	0.029	-	0.175	0.291				
	Alignment 1+2	M0	-19189.49	308	0.028	-	0.142	= d_N/d_S EPRV	0.46	1	4.98E-01	NS
		M1	-19189.26	309	0.028	-	0.148	0.130				
Positive selection	*Alignment Main BSUgV*	M7	-6576.11	82					0.00	2	1	NS
		M8	-6576.11	84								

LnL: Log-likelihood of the model
np: number of parameters of the model
d_N/d_S other: parameter for all branches except those of PRV and EPRV terminal branches
2dLnL: Twice the likelihood of the two compared models
df: number of degree of freedom
Significant at the 1% level, * at the 0.1% level

Table 4: d_N/d_S ratio, evolution rates estimates and likelihood ratio tests of PRV and EPRV terminal branches

Parameter	Dataset	Model	LnL	np	Branch label:			2dLnL	df	p	
					internal	PRV	EPRV				
r	*Alignment 1*	M0	-10176.41	77	1	0.942	=rPRV	7.91	1	4.91E-03	***
		M1	-10172.45	78	1	1.090	0.702				
	Alignment 2	M0	-14202.95	87	1	0.975	=rPRV	20.00	1	7.75E-06	***
		M1	-14192.95	88	1	1.338	0.739				
	Alignment 1+2	M0	-20059.24	155	1	0.940	=rPRV	13.23	1	2.76E-04	***
		M1	-20052.63	156	1	1.168	0.795				
d_N/d_S	*Alignment 1*	M0	-9620.40	152	0.042	0.039	= d_N/d_S PRV	19.04	1	1.28E-05	***
		M1	-9610.88	153	0.040	0.018	0.071				
	Alignment 2	M0	-13575.87	172	0.038	0.115	= d_N/d_S PRV	74.43	1	0	***
		M1	-13538.66	173	0.037	0.003	0.165				
	Alignment 1+2	M0	-19260.56	308	0.041	0.089	= d_N/d_S PRV	91.66	1	0	***
		M1	-19214.73	309	0.039	0.015	0.133				

LnL: Log-likelihood of the model
np: number of parameters of the model
Branch label = other: all branches except from PRV and EPRV terminal branches
2dLnL: Twice the likelihood of the two compared models
df: number of degree of freedom
***Significant at the 0.1% level

Furthermore, terminal branches of EPRVs evolved under a much lower selective pressure than terminal branches of PRVs. The d_N/d_S ratio of EPRVs was 4 and 50 times higher than those of PRV in group 1 and 2, respectively (Table 4). These results were similar to those of BSUgV sequences compared to PRV sequences (Table 3), confirming the probable endogenous nature of some BSUgV sequences.

Discussion

Definition of an appropriate BSV phylogeny

All phylogenies published so far focused either on episomal particles or on endogenous BSV. Our work combined for the first time all available episomal and integrated badnavirus sequences, thus contributing to the diversity of this viral genus. The maximum likelihood inference framework enabled us to reconstruct robust phylogenies confirmed by a Bayesian approach. The genetic diversity and phylogenetic relationships of BSV and badnaviruses was used to understand the evolution of the BSV species and to define an appropriate viral taxonomy.

BSV phylogeny obtained herein confirmed the existence of three distinct groups of BSV as suggested by Harper et al., 2005. Surprisingly, BSUgV sequences of groups 2 and 3 are phylogenetically more distantly related to the known BSV group 1 infecting banana, and more closely related to viruses that do not naturally infect banana such as SCBMV, ComYMV, CSSV and CMBV. This result clearly showed that BSV is polyphyletic. Analyses of sequences of full-length viral genomes of BSUgV group 2 and 3 would help to confirm that these sequences truly correspond to episomal viruses. Since the virus classification aims at reflecting the true phylogenetic relationships between species, a new taxonomic nomenclature of BSVs including the three groups should probably be established to avoid a polyphyletic classification.

Banana is a monocotyledon and the unique natural host of BSV species belonging to the three phylogenetic groups. KTSV, SCBMV, ComYMV, CSSV and CMBV are viruses that all clustered close to BSV, but these viral species are not able to infect banana under natural conditions. SCBMV infects sugarcane (monocotyledon) and CSSV is found in cocoa (dicotyledon). The phylogeny inferred herein suggested that in addition to BSV, banana is the ancestral host plant of the five above-mentioned badnaviruses and that at least five independent host plant shifts occurred during

badnaviral evolution. Such host changes can occur when two host plant species are colonized by the same viral insect vector (Harper et al. 2005; Jones 2000). In many tropical countries, banana is often grown near sugarcane and the mealybug *Planoccocus citri* - a vector of BSV - feeds on both plants (Jones 2000). Likewise, *P. citri* also feeds on citrus (Ben-Dov 1994), suggesting that this mealybug species contributed to the host change of SCBMV and CMBV. Interestingly, different isolates of SCBV from sugarcane are still able to infect banana when agroinoculated (Bouhida et al. 1993), suggesting either SCBV is a generalist or the host change was relatively recent. Nevertheless, this hypothesis needs to be validated by a full-length phylogeny of the badnaviruses and by obtaining the same topology than the one observed in this study. Likewise, further studies are needed to confirm that recombination does not alter the phylogeny inferred from the RT/RNaseH region used herein, however recombination between BSV species was not detected in nucleotidic alignments of BSV genomes (E. Muller, personal communication). If this topology and the host shifts directions are confirmed, a direct perspective of this result would be to search for adaptive mutations among viruses that underwent a host shift, in comparison to viruses of the same phylogenetic group that kept infecting banana only.

Harper et al., (2005) published sequences from Ugandan epidemics assumed to correspond to episomal viruses. The authors used IC-PCR and direct binding degenerate PCR to detect episomal viruses only. However, these methods often lead to a co-amplification of the virus but also of EPRVs from residual *Musa* genomic DNA. A systematic monitoring of EPRV contamination is required to insure that only episomal viruses are amplified (Le Provost et al. 2006). Our comparison of the evolution parameters of a set of sequences was very powerful to identify EPRVs among a set of episomal viral sequences, such as the BSUgV sequences. However, experimental verifications are required to assess, without the shadow of a doubt the real status of these sequences. BSUgAV sequences should be the first ones to be investigated, as they correspond to a new species in group 1, and where all known BSV isolates clustered. We are confident with the category assignment of the other sequences used in this paper because the PRV sequences originate from full-length

circular genomes and EPRV sequences derived from PCR performed on total DNA from virus free plant material.

BSV integration is a frequent and recent phenomenon

Combination of endogenous and episomal sequences in the same phylogeny provided evidence of large-scaled integrations of BSV in the genome of three *Musa* species. Two of them, *M. acuminata* and *M. balbisiana* were used in the domestication process of bananas. The majority of EPRVs diverging from a single integration event was restricted to a single *Musa* species, and was found in several genotypes of *M. balbisiana* and several subspecies of *M. acuminata* (Figure 3). This result suggested that integration events mainly occurred before the diversification of each host species, most likely before the domestication of bananas which is estimated to begin in the mid-Holocene around 6500 cal. years BP (Denham et al. 2003). Furthermore, in almost all cases, *i.e.* for 24 out of 27 integrations events, each phylogenetic group of EPRV diverging from a single integration event was found in a single host species. This indicated that the majority of integration events probably occurred after the speciation of the three *Musa* species, *i.e.* 4.6 My old for *M. acuminata* and *M. balbisiana* estimated from molecular evolution of zingiberales calibrated with paleontological data (Lescot et al. 2008).

It has been established that evolutionary data of pathogens provides a source of information about ancient host-parasite interactions (Nieberding and Olivieri 2007). If EPRVs are fossils of viral infections occurring in wild plant populations, they should reflect ancient BSV-banana interactions. Our study confirmed at a large phylogenetic level, that BSV species of group 1 mainly integrated the B genome (Gayral et al. 2008; Geering et al. 2001); Iskra-Caruana, unpublished data). It was first suggested that the diversity and actual distribution of EPRVs in *Musa* genomes could mirror the ancient geographical distribution of invasive PRVs and susceptible *Musa* host populations (Geering et al. 2005a). The *Musa balbisiana* species originates from South Asia: South India, Burma, North Philippines and New Guinea (Simmonds 1962), all possible locations for an ancient infection of BSV group 1 or their direct ancestors. Furthermore, previous work showed that several wild diploid *Musa balbisiana* plants are resistant to any multiplication of either endogenous or episomal

form of BSV group 1 (Iskra Caruana M.L. et al. 2003; Lheureux F. 2002). These results highlight a fossil plant-pathogen interaction between BSV group 1 and *Musa balbisiana* that lead to resistance of *M. balbisiana* against BSV group 1. The hypothesis that EPRV plays a role in plant resistance to the cognate virus has been often suggested in the literature (Hull et al. 2000; Mette et al. 2002; Staginnus and Richert-Pöggeler 2006) but direct experimental evidence is still missing. Gene silencing could play a role in such resistance mechanisms, because EPRV transcription and subsequent siRNA production occurred in *Petunia* and *Solanum* species (Noreen et al. 2007; Staginnus et al. 2007). A further notable characteristic of this group is that it is most likely represented solely by integrated sequences. Only eight BSUgV sequences belong to this group but were shown to be probably EPRVs (see above). These findings would mean that a lineage of episomal viruses existed in group 2 and that they were very successful in infecting plants, as suggested by the great diversity of this group, but were subsequently extinct. One other possibility is that episomal viruses of group 2 still exist but were not described yet. A better knowledge of the genetic diversity of BSV in other *Musa* genotypes and species, in other potential host and in different locations is needed to confirm this question.

BSV group 3 is also remarkable since no EPRV related to this group was found in the banana genome so far. Very few studies investigated on integrations of SCBV, a sister taxa of the BSV group 3, and there is no evidence of the existence of SCBV EPRVs in the sugarcane host genome (Geijskes et al. 2004); Muller E., pers. comm.). The lack of viral integration of group 3 remains unexplained. One hypothesis would be that their original host plants, with which these viruses could have coevolved and integrated before a host shift, have not been yet studied. It thus becomes necessary to screen the largest diversity within the genus *Musa* in order to find the putative ancestral hosts of group 3. Furthermore, the ongoing *Musa* sequencing project aims to sequence *Musa acuminata* ssp. *malaccencis* (cv. Pahang) will certainly help to determine which kind of BSV species integrated the A genome and if there are infectious EPRVs.

The large diversity of integrated badnaviruses found in the *Musa* genome is a consequence of multiple and recent independent integration events. The results obtained in our large scale study are in agreement with previous experimental work using limited BSV species. Fluorescent in situ hybridization (FISH) experiments and

fingerprint of *Musa* bacterial artificial chromosome (BAC) libraries showed that the *Musa* B genome harbors only a limited number of EPRV copies (Gayral et al. 2008; Harper et al. 1999; Iskra-Caruana, unpublished data). Conversely, the situation is totally different in other plants harboring EPRVs such as *Petunia* sp. and several *Solanaceae* species. Few initial EPRVs underwent large-scale amplification (hundreds to thousands copies) leading to the colonization of the whole host genome (Gregor et al. 2004; Mette et al. 2002; Richert- Pöggeler et al. 2003; Staginnus et al. 2007). This difference between banana and other plants could be related to the fact that BSV integrated the *Musa* genome more recently, and subsequent BSV EPRV amplification may not have occurred so far. Recent integration events could also explain why the viral ORFs of at least four EPRVs in *Musa* (BSOLV in cv. Obino L'Ewai and in cv. PKW, BSImV and BSGFV in cv. PKW) are conserved, and thus prone to induce infection after activation.

Detection of infectious EPRVs

The comparison of evolutionary parameters between PRV and EPRV sequences yielded unexpected results. Translation of the 99 EPRV sequences used in this study confirmed that 43 of them showed signs of pseudogenization, such as a substitution or an indel leading to a stop in frame or a premature stop codon, respectively. Furthermore, we observed a strong relaxation of the selective constraints of EPRVs (d_N/d_S = 0.13) compared to those of PRVs (d_N/d_S = 0.01). Despite evidence of pseudogenisation and a relaxation of the selective constraints, the d_N/d_S for EPRVs was still = 0.13, suggesting strong functional constraints on EPRV branches.

The second unexpected data resulted from the comparison of relative rates of evolution between PRV and EPRV sequences. Because very few studies on molecular evolution of plant pararetroviruses exist so far, we used data available for other viruses. The viral replication enzyme is an important determinant of evolutionary changes in viruses (Duffy et al. 2008), mostly because RdDps do not have proofreading capabilities and are therefore more error-prone than DNA polymerases (Flint et al. 2003). Substitution rate of viruses replicating with the same polymerase than do plant pararetroviruses (RNA-dependant DNA polymerases, RdDps or Reverse transcriptases, RTs) were therefore chosen for proxies of the substitution rate

of plant pararetroviruses. A substitution rate of 10^{-4}-10^{-5} substitutions per site per year (subs/site/year) was observed for *Hepatitis B virus* (*Hepadnaviridae*, animal pararetroviruses) (Zhou and Holmes 2007), a dsDNA virus closely related to the *Metaviridae* group that encompasses *Caulimoviridae*, Ty3/Gypsy retroelements and the DIRS group (Malik and Eickbush 2001). In addition, the substitution rate of retroviruses, which also use a RdDp and have a ssRNA genome, ranges 10^{-3}-10^{-6} subs/site/year (Hanada et al. 2004; Jenkins et al. 2002). We therefore hypothesized that the substitution rate of plant pararetroviruses also ranges 10^{-3}-10^{-6}. Likewise, an estimation of the synonymous substitution rate (d_S) of the *Musa* genome was 4.5 10^{-9} subs/site/year (Lescot et al. 2008), which reflected the neutral evolution rate of the host plant. Based on these estimations of substitution rates of episomal virus and host genome, at least three order of magnitude was expected between evolution rates of PRVs and EPRVs branches, but the difference was finally only a factor 1.5 (Table 4).

Two non exclusive hypotheses may explain these results. The first hypothesis assumes that some of the branches labeled as EPRV correspond to primarily episomal viruses that subsequently integrated the *Musa* genome. Estimations of evolutionary parameters in these mixed branches are therefore mean rates of PRVs and EPRVs. In comparison to expected rates for true EPRV branches, their d_N/d_S ratio should have decreased due to the selective pressures acting on PRVs, and their evolution rate should have increased due to the fast evolution of PRVs. One corollary of this hypothesis is that BSV integrations are recent since they occurred in the terminal branches of the phylogenetic tree.

The second hypothesis assumes that several branches labeled as episomal are not true episomal viruses. They could have had in their history an integrated past for a sufficient period of time between two episomal steps. These terminal branches would therefore correspond to both EPRV and PRV. In comparison to expected rates for PRV branches that did not experience integration, their d_N/d_S ratio should have increased due to the absence of selective pressures acting during the EPRV period, and accordingly their evolution rate should have decreased as the same speed as neutral evolution of the *Musa* genome, i.e. at least three order of magnitude slower than episomal viruses. However, the bias of sampling branches that correspond to

both EPRVs and PRVs was minimized by estimating the evolutionary parameters from terminal branches only and by shortening these branches by the use of all available EPRV and PRV sequences in our phylogenies. As both hypotheses modify the evolutionary parameters in the same way, it is difficult to choose between the two.

Activation of EPRVs was demonstrated only for PVCV/Petunia, TVCV/Tobacco and for four integrated BSV species so far, and infectious EPRVs were always been considered as scarce. However, these results suggest that activation of recently integrated PRV sequences is not marginal and are much more frequent than previously thought. BSV EPRVs played a significant role in *Musa*-BSV parasitic interactions, and consequently contributed to the diversity and evolution of episomal BSV.

Conclusion

Although we found herein that most EPRVs became rapidly pseudogenes, some EPRVs kept the ability to induce infection. However, many questions regarding the presence and evolution of EPRVs in the plant genomes remain unanswered. How did episomal viral genome integrate and why did no proper mechanism of viral integration ever evolve in plant pararetroviruses? In addition, infectious EPRVs are fascinating to study because they underlie a unique mode of virus transmission in plants. Alternatively, EPRVs can also be considered as a totally new form of emerging virus. From an evolutionary perspective, our results suggest that endogenous BSV sequences participated to BSV diversity and evolution via integrations and EPRV activations. Investigating BSV evolution, insertion polymorphism between the host genomes and molecular evolution of infectious EPRVs through time would greatly help to understand the reasons why deleterious elements are kept in the host genomes.

Acknowledgements

We are grateful to Nathalie Laboureau and Serge Galzi for technical assistance. We thank the members of our team and Elisabeth Fournier, Eric Bazin and Nicolas Galtier for their helpful comments, and Philippe Rott for improving the manuscript.

The materials and methods were conducted according to the quality management procedures of the standard ISO 9001:2000. P. G. was supported by a PhD grant CIRAD-Région Languedoc Roussillon.

Literature cited

Ben-Dov Y. 1994. A systematic catalogue of the mealybugs of the world (*Insecta*: *Homoptera*: *Coccoidea*: *Pseudococcidae* and *Putoidae*) with data on geographical distribution, host plants, biology and economic importance. Intercept Publications, Ltd., Andover, England, 686 pp.

Bouhida M, Lockhart BE, Olszewski NE. 1993. An analysis of the complete sequence of a *Sugarcane bacilliform virus* genome infectious to banana and rice. J Gen Virol 74:15-22.

Denham TP, Haberle SG, Lentfer C, Fullagar R, Field J, Therin M, Porch N, Winsborough B. 2003. Origins of agriculture at Kuk Swamp in the highlands of New Guinea. Science 301(5630):189-193.

Duffy S, Shackelton LA, Holmes EC. 2008. Rates of evolutionary change in viruses: patterns and determinants. Nat Rev Genet 9(4):267-276.

Fauquet CM, Mayo MA, Maniloff J, Desselberger U, Ball LA. 2005. Virus Taxonomy, VIIIth Report of the ICTV: London: Elsevier/Academic Press.

Flint SJ, Enquist LW, Skalka AM. 2003. Principles of Virology: Molecular Biology, Pathogenesis, and Control of Animal Viruses, 2nd Edition: ASM Press. p 850.

Gaut BS, Wright SI, Rizzon C, Dvorak J, Anderson LK. 2007. Recombination: an underappreciated factor in the evolution of plant genomes. Nat Rev Genet 8(1):77-84.

Gawel NJ, Jarret, R.L. 1991. A modified CTAB DNA extraction procedure for *Musa* and *Ipomoea*. Plant Mol Biol Rep 9:262-266.

Gayral P, Noa-Carrazana J-C, Lescot M, Lheureux F, Lockhart BEL, Matsumoto T, Piffanelli P, Iskra-Caruana M-L. 2008. A single *Banana streak virus* integration event in the banana genome as the origin of infectious endogenous pararetrovirus. J Virol 82(13):6697-6710.

Geering ADW, McMichael LA, Dietzgen RG, Thomas JE. 2000. Genetic diversity among *Banana streak virus* isolates from Australia. Phytopathology 90(8):921-927.

Geering ADW, Olszewski NE, Dahal G, Thomas JE, Lockhart BEL. 2001. Analysis of the distribution and structure of integrated *Banana streak virus* DNA in a range of *Musa* cultivars. Mol Plant Pathol 2(4):207-213.

Geering ADW, Olszewski NE, Harper G, Lockhart BEL, Hull R, Thomas JE. 2005a. Banana contains a diverse array of endogenous badnaviruses. J Gen Virol 86:511-520.

Geering ADW, Pooggin MM, Olszewski NE, Lockhart BEL, Thomas JE. 2005b. Characterisation of *Banana streak Mysore virus* and evidence that its DNA is integrated in the B genome of cultivated *Musa*. Arch Virol 150(4):787-796.

Geijskes RJ, Braithwaite KS, Smith GR, Dale JL, Harding RM. 2004. Sugarcane bacilliform virus encapsidates genome concatamers and does not appear to

integrate into the Saccharum officinarum genome. Archives of Virology 149(4):791-798.
Gregor W, Mette MF, Staginnus C, Matzke MA, Matzke AJM. 2004. A distinct endogenous pararetrovirus family in *Nicotiana tomentosiformis*, a diploid progenitor of polyploid tobacco. Plant Physiol 134(3):1191-1199.
Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Systematic Biology 52(5):696-704.
Hall TA. 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucl Acids Symp Ser 41:95-98.
Hanada K, Suzuki Y, Gojobori T. 2004. A large variation in the rates of synonymous substitution for RNA viruses and its relationship to a diversity of viral infection and transmission modes. Mol Biol Evol 21(6):1074-1080.
Harper G, Hart D, Moult S, Hull R. 2004. *Banana streak virus* is very diverse in Uganda. Virus Res 100(1):51-56.
Harper G, Hart D, Moult S, Hull R, Geering A, Thomas J. 2005. The diversity of *Banana streak virus* isolates in Uganda. Arch Virol 150(12):2407-2420.
Harper G, Hull R. 1998. Cloning and sequence analysis of *Banana streak virus* DNA. Virus Genes 17(3):271-278.
Harper G, Osuji JO, Heslop-Harrison JSP, Hull R. 1999. Integration of *Banana streak badnavirus* into the *Musa* genome: molecular and cytogenetic evidence. Virology 255(2):207-213.
Hohn T, Richert- Pöggeler KR, Harper G, Schwarzacher T, Teo CH, Techeney PY, Iskra-Caruana ML, Hull R. 2008. Evolution of integrated plant viruses. In Virus Evolution. Springer, Heidelberg, M. Roosinck ed. In press.
Huelsenbeck JP, Ronquist F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17(8):754-755.
Hull R, Harper G, Lockhart B. 2000. Viral sequences integrated into plant genomes. Trends Plant Sci 5(9):362-365.
Iskra Caruana M.L., Lheureux F., Noa-Carrazana J.C., Piffanelli P., Carreel F., Jenny C., Laboureau N., Lockhart B.E.L. 2003. Unstable balance of relation between pararetrovirus and its host plant: the BSV-EPRV banana pathosystem. Abstracts : EMBO Workshop Genomic Approaches in Plant Virology, May 28-31. Keszthely, Hungary. p 8.
Jakowitsch J, Mette MF, van der Winden J, Matzke MA, Matzke AJM. 1999. Integrated pararetroviral sequences define a unique class of dispersed repetitive DNA in plants. PNAS 96(23):13241-13246.
Jaufeerally-Fakim Y, Khorugdharry A, Harper G. 2006. Genetic variants of *Banana streak virus* in Mauritius. Virus Res 115(1):91-98.
Jenkins GM, Rambaut A, Pybus OG, Holmes EC. 2002. Rates of molecular evolution in RNA viruses: a quantitative phylogenetic analysis. J Mol Evol 54(2):156-165.
Jones DR. 2000. Diseases of banana, abacá, and enset. Wallingford, Oxon, UK ; New York: CABI Pub. pp xv, 544.
Kunii M, Kanda M, Nagano H, Uyeda I, Kishima Y, Sano Y. 2004. Reconstruction of putative DNA virus from endogenous rice tungro bacilliform virus-like sequences in the rice genome: implications for integration and evolution. BMC Genomics 5(1):80.

Le Provost G, Iskra-Caruana ML, Acina I, Teycheney PY. 2006. Improved detection of episomal Banana streak viruses by multiplex immunocapture PCR. J Virol Methods 137(1):7-13.

Lescot M, Piffanelli P, Ciampi AY, Ruiz M, Blanc G, Leebens-Mack J, da Silva FR, Santos CM, D'Hont A, Garsmeur O, Vilarinhos AD, Kanamori H, Matsumoto T, Ronning CM, Cheung F, Haas BJ, Althoff R, Arbogast T, Hine E, Pappas GJ, Jr., Sasaki T, Souza MT, Jr., Miller RN, Glaszmann JC, Town CD. 2008. Insights into the *Musa* genome: syntenic relationships to rice and between *Musa* species. BMC Genomics 9(1):58.

Lheureux F, Laboureau N, Muller E, Lockhart BE, Iskra-Caruana ML. 2007. Molecular characterization of *Banana streak acuminata Vietnam virus* isolated from *Musa acuminata siamea* (banana cultivar). Arch Virol 152(7):1409-1416.

Lheureux F. 2002. Etude des mécanismes génétiques impliqués dans l'expression des séquences EPRVs pathogènes des Bananiers au cours de croisements génétiques interspécifiques [PhD]. Montpellier: Université Sciences et Techniques du Languedoc USTL.

Lockhart B, Jones D. 2000. Banana streak. In Diseases of Banana, Abaca and Enset, ed. DR Jones, pp. 263–74. Wallingford, UK CAB Int.

Lockhart BE, Menke J, Dahal G, Olszewski NE. 2000. Characterization and genomic analysis of *Tobacco vein clearing virus*, a plant pararetrovirus that is transmitted vertically and related to sequences integrated in the host genome. J Gen Virol 81:1579-1585.

Malik HS, Eickbush TH. 2001. Phylogenetic analysis of ribonuclease H domains suggests a late, chimeric origin of LTR retrotransposable elements and retroviruses. Genome Res 11(7):1187-1197.

Mette MF, Kanno T, Aufsatz W, Jakowitsch J, van der Winden J, Matzke MA, Matzke AJM. 2002. Endogenous viral sequences and their potential contribution to heritable virus resistance in plants. Embo J 21(3):461-469.

Ndowora T, Dahal G, LaFleur D, Harper G, Hull R, Olszewski NE, Lockhart B. 1999. Evidence that badnavirus infection in *Musa* can originate from integrated pararetroviral sequences. Virology 255(2):214-220.

Nieberding CM, Olivieri I. 2007. Parasites: proxies for host genealogy and ecology? Trends Ecol Evol 22(3):156-165.

Noreen F, Akbergenov R, Hohn T, Richert- Pöggeler KR. 2007. Distinct expression of endogenous *Petunia vein clearing virus* and the DNA transposon dTph1 in two *Petunia hybrida* lines is correlated with differences in histone modification and siRNA production. Plant Journal 50(2):219-229.

Pahalawatta V, Druffel K, Pappu H. 2008. A new and distinct species in the genus Caulimovirus exists as an endogenous plant pararetroviral sequence in its host, Dahlia variabilis. Virology 376(2):253-257.

Posada D, Buckley TR. 2004. Model selection and model averaging in phylogenetics: advantages of akaike information criterion and bayesian approaches over likelihood ratio tests. Syst Biol 53(5):793-808.

Posada D, Crandall KA. 1998. MODELTEST: testing the model of DNA substitution. Bioinformatics 14(9):817-818.

Richert- Pöggeler KR, Noreen F, Schwarzacher T, Harper G, Hohn T. 2003. Induction of infectious petunia vein clearing (pararetro) virus from endogenous provirus in petunia. Embo J 22(18):4836-4845.

Richert- Pöggeler KR, Shepherd RJ. 1997. *Petunia vein-clearing virus*: a plant pararetrovirus with the core sequences for an integrase function. Virology 236(1):137-146.

Schuermann D, Molinier J, Fritsch O, Hohn B. 2005. The dual nature of homologous recombination in plants. Trends Genet 21(3):172-181.

Simmonds NW, editor. 1962. The evolution of the bananas, Longmans Green (Ed). London. p 170.

Staginnus C, Gregor W, Mette MF, Teo CH, Borroto-Fernandez EG, Machado ML, Matzke M, Schwarzacher T. 2007. Endogenous pararetroviral sequences in tomato (*Solanum lycopersicum*) and related species. BMC Plant Biol 7:24.

Staginnus C, Richert- Pöggeler KR. 2006. Endogenous pararetroviruses: two-faced travelers in the plant genome. Trends Plant Sci 11(10):485-491.

Su L, Gao S, Huang Y, Ji C, Wang D, Ma Y, Fang R, Chen X. 2007. Complete genomic sequence of *Dracaena mottle virus*, a distinct badnavirus. Virus Genes 35(2):423-429.

Swofford DL. 2002. PAUP* 4: Phylogenetic Analysis Using Parsimony (and other methods). Sinauer, Sunderland, Massachusetts, USA.

Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucl Acids Res 22(22):4673-4680.

Xia X. 1999. DAMBE (Software Package for Data Analysis in Molecular Biology and Evolution). User Manual. Hong Kong: Department of Ecology and Biodiversity, University of Hong Kong.

Xia X, Xie Z. 2001. DAMBE: software package for data analysis in molecular biology and evolution. J Hered 92(4):371-373.

Xia X, Xie Z, Salemi M, Chen L, Wang Y. 2003. An index of substitution saturation and its application. Mol Phyl Evol 26(1):1-7.

Yang IC, Hafner GJ, Dale JL, Harding RM. 2003. Genomic characterisation of taro bacilliform virus. Arch Virol 148(5):937-949.

Yang Z. 1997. PAML: a programme package for phylogenetic analysis by maximum likelihood. Comp Appl Biosci 13:555-556.

Yang Z. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol Biol Evol 15:568–573.

Yang Z, Nielsen R. 1998. Synonymous and nonsynonymous rate variation in nuclear genes of mammals. J Mol Evol 46(4):409-418.

Yang Z, Yoder AD. 2003. Comparison of likelihood and Bayesian methods for estimating divergence times using multiple gene Loci and calibration points, with application to a radiation of cute-looking mouse lemur species. Syst Biol 52(5):705-716.

Yoder AD, Yang Z. 2000. Estimation of primate speciation dates using local molecular clocks. Mol Biol Evol 17(7):1081-1090.

Zhou Y, Holmes EC. 2007. Bayesian estimates of the evolutionary rate and age of hepatitis B virus. J Mol Evol 65(2):197-205.

2 Histoire évolutive des intégrations pathogènes du *Banana streak GF virus* et du *Banana streak Im virus* chez leur hôte : apports de la phylogénie moléculaire du genre *Musa*

2.1 Objectifs généraux

L'objectif de cette étude est de retracer l'histoire évolutive des EPRV infectieux de deux espèces de BSV : BSGFV et BSImV, depuis leur intégration dans le génome des bananiers.

Afin de préciser l'origine des intégrations, nous avons dans un premier temps établi une phylogénie de la diversité des espèces, des sous-espèces et des génotypes de bananiers, en nous focalisant notamment sur l'espèce *M. balbisiana* chez laquelle ces EPRV ont été décrits.

En nous basant sur la phylogénie des hôtes, nous avons ensuite déterminé les dates relatives d'intégration de ces virus dans le génome des bananiers par l'analyse du polymorphisme d'intégration. Nous avons également pu suivre l'évolution de la structure des EPRV au cours de la diversification des hôtes.

Finalement, la construction de phylogénies moléculaires nous a permis de retrouver, parmi les génotypes des virus libres BSGFV et BSImV, ceux qui ont probablement été à l'origine de chacune des intégrations dans le génome de *M. balbisiana.*

Ces travaux sont présentés ci-après sous la forme d'un article qui a été soumis à la revue *BMC Evolutionary Biology* :

P. Gayral, L. Blondin, O. Guidolin, F. Carreel, I. Hippolyte, X. Perrier and M.-L. Iskra-Caruana. Evolutionary history of infectious endogenous banana streak viruses and their host banana (*Musa* sp.).

2.2 Article 4 : "Evolutionary history of infectious endogenous banana streak viruses and their host banana (*Musa* sp.)"

Evolutionary history of infectious endogenous banana streak viruses and their host banana (*Musa* sp.)

Philippe Gayral[1], Laurence Blondin[1], Olivier Guidolin[1], Françoise Carreel[1], Isabelle Hippolyte[2], Xavier Perrier[2] and Marie-Line Iskra-Caruana[1]

[1]CIRAD BIOS, UMR Biologie et Génétique des Interactions Plante-Parasite (BGPI), TA A-54 / K, Campus international de Baillarguet, 34398 Montpellier Cedex 5, France.

[2]CIRAD BIOS, UPR Amélioration génétique d'espèces à multiplication végétative, TA A-75 / 02, Avenue Agropolis, 34398 Montpellier Cedex 5 France

Corresponding author: M.-L. Iskra-Caruana

Tel: (+33) 4 99 62 48 13

Fax : (+33) 4 99 62 48 08

e-mail: marie-line.caruana@cirad.fr

Abstract

Endogenous plant pararetroviruses (EPRVs) are viral sequences of the family *Caulimoviridae* integrated into the nuclear genome of numerous plant species. *Banana streak virus*es (BSVs) have the amazing feature to result from either episomal viruses or infectious EPRVs. Although such EPRV probably arose from accidental events, they constitute an extreme case of parasitism and an original strategy of horizontal transmission of plant viruses. In this study, we investigated the early evolutionary stages of infectious EPRVs of two BSV species, Goldfinger - BSGFV and Imové - BSImV in relation to the host banana (*Musa* genus). We first analyzed their distribution among the genus *Musa* by studying their insertion polymorphism and structure evolution using 13 PCR markers. To propose scenarios of the evolution of BSV integrations, we inferred a phylogeny of the banana carrying these EPRVs from 2.1 kbp of the chloroplast genome using *matK* gene and *trnL-trnF* regions, and from the nuclear genome using 19 microsatellite loci. Finally, we established the origins of integrated BSGFV and BSImV by studying phylogenetic relationships between ERPV sequences and their non integrated counterparts.

Key-words

Integrated viral sequences, Badnavirus integration, *Banana streak virus* (BSV), *Musa balbisiana*, Viral evolution, Chloroplast phylogeny, Microsatellites phylogeny.

Introduction

Caulimoviridae (plant pararetroviruses) replicate by reverse transcription from transcribed pregenomic RNA (Hull and Covey 1995) and are phylogenetically close to *Metaviridae* (Ty3-Gypsy elements) (Malik and Eickbush 2001). All caulimoviruses have a non-covalently closed circular double-stranded DNA genome of 7-8 kbp (Hull 1999). They exist as episomal (i.e. exogenous) viruses and as integrated (i.e. endogenous) sequences (endogenous pararetrovirus - EPRV) in the host plant genome, although the viral replication cycle excludes an integration step. Both types are infectious and induce a viral multiplication in plants. Up to date, EPRVs were described in 9 distantly related mono- and dicotyledon plant families. Each originated from independent integrations events from 5 out of the 6 genera of the *Caulimoviridae* family (Pahalawatta et al. 2008; Staginnus and Richert-Poggeler 2006; Su et al. 2007). EPRVs are not systematically eliminated from plant genome According to a process of endogenization; viral sequences first integrated the germinal cells to become part of the plant genome. Following, EPRVs were fixed in plant populations by evolutionary forces such as natural selection and/or genetic drift (Hohn et al. 2008; Hull and Covey 1995). Integration mechanism is supposed to occur via illegitimate recombination between plant and viral genomes (Staginnus and Richert-Poggeler 2006).

Despite their accidental origin, the presence of EPRV has three major consequences for host plants. First, in petunia (*Petunia* sp.) and tobacco (*Nicotiana* sp.), EPRV reached several hundreds to thousands copies, respectively (Gregor et al. 2004; Jakowitsch et al. 1999; Richert-Poggeler et al. 2003), mainly found in heterochromatin. This amplification was probably the result of integration mediated by transposable elements found embedded or closed to EPRVs (Matzke et al. 2004; Staginnus and Richert-Poggeler 2006). EPRVs may therefore contribute to genome size modification, may induce changes in the methylation status of the host genome, and may also be genomic reorganizers by inducing chromosomal rearrangements (Hohn et al. 2008), such as transposable elements (Bennetzen 2000; Kidwell and Lisch 2000).

Then, EPRV is supposed to serve host functions. A low level of EPRV transcription and subsequent siRNA production was observed in petunia (*Petunia hybrida*) and tomato (*Solanum lycopersicum*) (Noreen et al. 2007; Staginnus et al. 2007). It is assumed that EPRVs-induced homology-dependant gene silencing targeted the counterpart exogenous viruses: *Petunia vein clearing virus* (PVCV - genus Petuvirus) in petunia, and *Tobacco vein clearing virus* (TVCV - genus Cavemovirus) in tomato. This mechanism is also suspected for BSV (genus *badnavirus*) in wild diploid *Musa balbisiana*, resistant to both activation of infectious EPRVs and BSV inoculation by insect vector (Hohn et al. 2008). Resistance mechanisms against episomal and endogenous forms of *Caulimoviridae* are seen as plants counter adaptations in response to the presence of infectious EPRVs in their genomes.

Finally EPRVs are able to be infectious by releasing a functional full-length viral genome, from which a viral multiplication and a plant infection occurs. To date, infectious EPRVs are observed in 3 pathosystems: BSV - banana, PVCV - petunia, and TVCV - tobacco. New genome combinations through interspecific hybridization trigger activation in all pathosystems (Lheureux et al. 2003; Lockhart et al. 2000; Richert-Poggeler et al. 2003), other stress conditions such as wounding trigger activation of PCVC in Petunia (Richert-Poggeler et al. 2003), and in vitro culture for BSV in banana (Dallot et al. 2001). In banana, the triploid interspecific hybrids (AAB genome) between diploids *Musa acuminata* cv IDN1104x (AAAA genome) and *Musa balbisiana* cv. Pisang Klutuk Wulung (PKW) (BB genome) is a good model to study activation of BSV EPRV. Two molecular mechanisms are suspected to enable activation of infectious EPRV: an homologous recombination between repeat regions surrounding EPRVs resulting in the excision of a circular viral genome (Ndowora et al. 1999); and the transcription of EPRVs leading to a viral pregenomic RNA (Noreen et al. 2007; Richert-Poggeler et al. 2003).

The presence of infectious EPRVs also impacts the evolution of the virus itself. Infectious EPRVs are an extreme case of parasitism and use a newly described strategy of viral transmission among plant viruses. However, the endogenous state may be disadvantageous for the virus, since EPRV accumulate deleterious mutations with time and rapidly become defective.

Banana streak virus (BSV) naturally infects banana species (*Musa* sp.) and is found in the entire banana producing area (Lockhart and Jones 2000). The use of the error-prone viral reverse-transcriptase would in part account for the genetic diversity and species richness observed in some genera (Duffy et al. 2008), such as in badnavirus (Fargette et al. 2006). BSV is indeed the generic name of several species showing up to 30 % nucleotidic divergence but provoking the same disease on banana. Based on RT-RNAseH region of ORFIII, three major phylogenetic groups (BSV-1 to BSV-3) were observed (Harper et al. 2005). *Banana streak Imové virus* (BSImV) and *Banana streak Goldfinger virus* (BSGFV) belong to Group BSV-1, which also encompass at least five other fully described BSV species. Groups BSV-2 and BSV-3 each contain dozens of putative species that remain to be fully characterized with whole genome data and electronic microscopy (Bousalem et al. 2008; Gayral et al, unpublished data; Harper et al. 2005).

BSV-banana appeared a good model system to investigate the evolution of infectious EPRVs. First, BSV phylogeny was actively investigated, and a good picture of its genetic diversity and of the BSV integrations existing in the *Musa* genome are now available (Bousalem et al. 2008; Gayral al, unpublished data, Geering et al. 2005). Then, BSImV and BSGFV EPRV are present in a single copy in the *Musa balbisiana* cv. PKW genome (Gayral et al. 2008, Gayral et al., unpublished data), unlike EPRV in *Solanaceae* reaching several hundreds to thousands copies per genome (Gregor et al. 2004; Jakowitsch et al. 1999; Richert-Poggeler et al. 2003; Staginnus et al. 2007). Furthermore, the biology of infectious integrants was investigated and hypothesis of mechanisms of activation (Gayral al, unpublished data), genetic and environmental factors triggering EPRV activation are better understand in banana (Dallot et al. 2001; Lheureux et al. 2003). However, very little is known regarding to their evolutionary history that may explain the nowadays presence and distribution of infectious EPRVs in the *Musa* genus. Furthermore, the phylogeny of the host species (*Musa balbisiana*) is largely unknown. The few studies concerning the genetic diversity of *Musa balbisiana* focused on region-specific sampling rather than on exhaustive genetic diversity (Ge et al. 2005; Uma et al. 2005). Consequently, no comparative studies

connecting EPRV distribution to the host phylogeny could be made. Information on the evolution of this unusual biological model is helpful to understand the phenomenon of *Caulimoviridae* integration in other plants, and to propose mechanisms explaining the presence of infectious EPRVs in plants.

In this study, we aimed to investigate the early evolutionary stages of infectious EPRVs following their integration, in relation to their host bananas (*Musa* sp.). We focused on infectious integrations of two *Banana streak virus* species: Imové - BSImV and Golfinger - BSGFV. Each integrant is present in a single copy in the genome of *Musa balbisiana* cv. 'Pisang Klutuk Wulung' (PKW). They were studied for their ability to restitute infectious viruses in interspecific hybrids between cv. PKW, and *Musa acuminata* cv. 'IDN 110 4x' (Iskra Caruana M.L. et al. 2003; Lheureux et al. 2003). BSImV EPRV locus is homozygous in the cv. PKW genome, and BSGFV EPRV locus carried two alleles, only one being infectious.

The first question we addressed was to examine the recent evolutionary fate following integration in the host genome, in terms of integration date, of conservation, degradation or loss of EPRV. We answered this question by analyzing the insertion polymorphism and the evolution of the structure of these EPRVs among 61 banana accessions representative of the diversity of the genus *Musa*, using 13 PCR markers specific of BSImV and BSGFV. The second question was to understand the evolution of the interaction between hosts and EPRVs, and to determine if the two infectious endogenous BSV species underwent the same evolutionary history. To do so, we analyzed the distribution of EPRVs among 32 accessions of the *Musa* genus in line with the host phylogeny inferred from 19 microsatellites loci and from molecular phylogeny based on *matK* and *trnL-trnF* chloroplast regions. Scenarios of BSGFV EPRV evolution were also based upon the estimation of the age of the integration and the timing of evolution from sequence divergence. Finally, we determined the origins and nearest ancestors of the corresponding EPRV sequences using phylogenies of all available published episomal BSImV and BSGFV sequences.

Materials and Methods

Plant materials and DNA extraction

The plant material listed in the Table 1 was chosen to represent the four sections of the genus *Musa*. Accessions of the wild *Musa balbisiana* (BB genome) were predominantly sampled. *Ensete ventricosum*, a species from the only other genus in the family *Musaceae* was included as an outgroup taxon. Fresh leaf samples were kindly supplied from the CIRAD collection in Guadeloupe and from the International Institute of Tropical Agriculture (IITA) in Cameroun; dried leaf samples and plantlets from *in-vitro* culture were supplied from the International Transit Center (ITC) in Leuven, Belgium.

Total genomic DNA was extracted from leaf tissue by the method of Gawel and Jarret (1991) (Gawel 1991). The quality of DNA was visually estimated after migration of 5 µl of DNA in a 0.8 % agarose gel, staining with ethidium bromide, and visualizing under UV light.

Table 1: *Musa* accessions and screen of EPRV of plants used in this study

Section	Species/ Group/ Hybrid	Phylogeny[c]	Subspecies	Genome[c]	Name[d]	Abreviations[d]	Accession nb.[b]	EPRV BSImV[fg]	EPRV BSGFV[fh]
Ensete[a]	*E. glaucum* (Roxb) Cheesm.	Cp	-	-	Vudu Vudu	-	ITC0775	-----	--------
Australimusa	*M. textilis* Née	-	-	TT	-	-	NEU0001	-----	--------
Callimusa	*M. beccarii* Simmonds	-	-	TT	-	-	NEU0005	-----	--------
Callimusa	*M. coccinea* Andr.	-	-	-	-	-	NEU0003	-----	--------
Rhodochlamys	*M. laterita* E. E. Cheesm.	Cp/Ms	-	-	-	-	NEU0008	-----	--------
Rhodochlamys	*M. ornata Roxb.*	Cp/Ms	-	-	-	-	NEU0007	-----	--------
Rhodochlamys	*M. mannii* H.Wendl	Cp/Ms	-	-	-	-	NEU0011	-----	--------
Rhodochlamys	*M. velutina* H.Wendl & Drude	Cp/Ms	-	-	-	-	ITC0638	-----	--------
Rhodochlamys	*M. sanguinea*	Cp/Ms	-	-	-	-	NEU0010	-----	--------
Eumusa	*M. basjoo*	Cp/Ms	-	-	-	-	NEU0060	-----	--------
Eumusa	*M. boman*	Cp/Ms	-	-	-	-	ITC1026	-----	++-+--++
Eumusa	*M. itinerans* E. E. cheesm.	Cp/Ms	-	-	-	-	-	-----	--------
Eumusa	*M. schizocarpa* Simmonds	-	-	SS	M. Schizocarpa N1	-	ITC0599	-----	--------
	M. schizocarpa Simmonds	-	-	SS	Musa schizocarpa	-	ITC0852	-----	--------
	M. schizocarpa Simmonds	-	-	SS	Musa schizocarpa	-	ITC0890	-----	--------
	M. schizocarpa Simmonds	-	-	SS	Musa schizocarpa	-	ITC0926	-----	--------
	M. schizocarpa Simmonds	-	-	SS	Musa schizocarpa	-	ITC1002	-----	--------
	M. schizocarpa Simmonds	-	-	SS	Musa schizocarpa	-	ITC1024	-----	--------
	M. schizocarpa Simmonds	Cp/Ms	-	SS	Musa schizocarpa	-	ITC0846	-----	--------
	M. schizocarpa Simmonds	-	-	SS	Musa schizocarpa	-	ITC0856	-----	--------
Eumusa	*M. acuminata* Colla	-	ssp. *errans* Blanco	AAw	Agutay		NEU0033	-----	--------
	M. acuminata Colla	-	ssp. *microcarpa* Simmonds	AAw	Bornéo		NEU0028	-----	--------
	M. acuminata Colla	-	ssp. *truncata* (Ridl.) Shepherd	AAw	Truncata		NEU0027	-----	--------
	M. acuminata Colla	-	ssp. *siamea* Simmonds	AAw	Khae (Phrae)		NEU0025	-----	--------
	M. acuminata Colla	-	ssp. *siamea* Simmonds	AAw	Pa (Rayong)		NEU0024	-----	--------
	M. acuminata Colla	Cp/Ms	ssp. *burmanicoides* (De Langhe & Devreux)	AAw	Calcutta 4	BUR	NEU0017	-----	--------
	M. acuminata Colla	-	ssp. *burmanica* Simmonds	AAw	Long Tavoy		NEU0016	-----	--------

	M. acuminata Colla	Cp/Ms	ssp. *malaccencis* Simmonds	AAw	Pahang	MAL	NEU0013	-----	--------
	M. acuminata Colla	-	ssp. *malaccencis* Simmonds	AAw	Malaccensis		NEU0012	-----	--------
	M. acuminata Colla	-	ssp. *malaccencis* Simmonds	AAw	P. cici (bresil)		NEU0015	-----	--------
	M. acuminata Colla	Cp/Ms	ssp. *zebrina nom. nud.*	AAw	Zebrina	ZEB	NEU0029	-----	--------
	M. acuminata Colla	-	ssp. *banksii* (F. Muell) Simmonds	AAw	Madang		ITC0254	-----	--------
	M. acuminata Colla	-	-	AAw	Pa (Songkhla)		NEU0043	-----	--------
	M. acuminata Colla	-	-	AAw	THA018		NEU0034	-----	--------
	M. acuminata Colla	-	-	AAw	Rung Hoa Xoan		ITC1432	-----	--------
	M. acuminata Colla	-	-	AAcv	Sa		NEU0149	-----	--------
	M. acuminata Colla	-	-	AAcv	Idn 110 d		NEU0137	-----	--------
	Banksii	-	-	AAcv	Wikago		NEU0102	-----	--------
	Desert banana	-	-	AAcv	Ta		NEU0096	-----	--------
	Rio	-	-	AAAcv	Leite		NEU0226	-----	--------
	Cavendish	Cp	-	AAAcv	Grande Naine	CAV	NEU0172	-----	--------
Eumusa	*M. balbisiana* Colla	Cp/Ms	-	BBw	Pisang Klutuk Wulung	M.b PKW	NEU0054	+++++	++++++++
	M. balbisiana Colla	Cp/Ms	-	BBw	Pisang Klutuk	M.b PKL	NEU0056	+++++	++++++++
	M. balbisiana Colla	Cp/Ms	-	BBw	Pisang Batu	M.b PBA	NEU0055	+ - +++	++++++++
	M. balbisiana Colla	Cp/Ms	-	BBw	Klue Tani	M.b KTA	NEU0053	+++++	++++++++
	M. balbisiana Colla	Cp/Ms	-	BBw	Honduras	M.b HDN	NEU0049	-----	++++++++
	M. balbisiana Colla	Cp/Ms	-	BBw	Balbisiana (10852)	M.b 852	ITC0094/ ONN0149	+++++	++++++++
	M. balbisiana Colla	Cp/Ms	-	BBw	-	M.b 545	ITC0545	+++++	++++++++
	M. balbisiana Colla	Cp/Ms	-	BBw	Cameroun	M.b CAM	NEU0050	++++-	+++++ - ++
	M. balbisiana Colla	Cp/Ms	-	BBw	Lal Velchi	M.b LVE	NEU0051	-----	+++++ - - +
	M. balbisiana Colla	Cp/Ms	-	BBw	Singapuri	M.b SIN	NEU0052	++ - - +	+++++ - ++
	M. balbisiana Colla	Cp/Ms	-	BBw	Los Banos	M.b LBA	ONN0151	+++++	+++++ - ++
	M. balbisiana Colla	Cp/Ms	-	BBw	Montpellier	M.b MPL	ONN0152	+++++	+++++ - ++
	M. balbisiana Colla	Cp/Ms	-	BBw	I 63	M.b I 63	ONN0154	-----	+++++ - ++
	M. balbisiana Colla	Cp/Ms	-	BBw	Eti Kehel	M.b EKE	ITC0271	++ - - -	+++++ - ++
	M. balbisiana Colla	Cp/Ms	-	BBw	-	M.b HDN-211	ITC0211	+++++	+++++ - ++
	M. balbisiana Colla	Cp/Ms	-	BBw	-	M.b I63-080	ITC0080	-----	+++++ - ++

	M. balbisiana Colla	Cp/Ms	-	BBw	-	M.b LBA-342	ITC0342	+++++	+++++ - ++
	M. balbisiana Colla	Cp/Ms	-	BBw	-	M.b 626	ITC0626	+++++	+++++ - ++
	M. balbisiana Colla	Cp/Ms	-	BBw	-	M.b 1016	ITC1016	++++ -	++++++++
	Hyb. *M. balbisiana*	Cp/Ms	-	BB	Butohan	M.b BUT	ITC0565/ NEU0057	+++++	++++++++

[a] *E. ventricosum* is in genus Ensete
[b] Used in this study for microsatellites (Ms) and chloroplast molecular phylogeny (Cp)
[c] w = wild type, cv = cultivar
[d] Name and Abreviation = if different from species name
[e] Collections : ITC, INIBAP Transit Center; NEU, CIRAD-Neufchâteau; ONNE, IITA
[f] '+' = presence of PCR amplification at the expected size; '-' = absence of amplification
[g] order of PCR markers for BSImV EPRV: Musa/F2, F1/F3, F3/F4, F4/F5, F5/Musa
[h] order of PCR markers for BSGFV EPRV: VM1, VV1, VV2, VV3, VV4, VV5, VV6, VM2

Nucleotide sequences

Two chloroplast loci were sequenced for the phylogenetic analysis of *Musa* genus. A newly designed pair of primers amplified 1,250 bp of *trnK* intron and partial *matK* gene, using primers MatKHB1Musa (5′ ATGGAAGAATTACAAGGATATTTAG 3′) and MatK1326RMusa (5′ AGCACACGAAAGTCGAAGT 3′) with 1.5 mM $MgCl_2$. These new primers were adapted from primers MatKHB1 (http://www.plantbio.ohiou.edu/epb/faculty/faculty/heb.htm) and MatK1326R (Cuenoud et al. 2002), respectively, after alignments with *matK* sequences from *Musa acuminata* (Gb: EU016987), *M. basjoo* (Emb: AJ581437), *M. beccari* (Gb AF434869) and *M. rosea* (Emb: AM114725). Then, the primers *trnL-F* C (5′ CGAAATCGGTAGACGCTACG 3′) and *trnF* F (5′ ATTTGAACTGGTGACACGAG 3′) (Taberlet et al. 1991) amplified 800-900 bp of the *trnL* intron and *trnL-trnF* intergenic spacer and were used with 3 mM $MgCl_2$.

Phylogeny of BSGFV was performed from sequences amplified using primers PhyloEPRV-Gf-F (5′ CAGCTCCAGGAGATTGGAAA 3′) and PhyloEPRV-Gf-R (5′ GGAGGAATCTATCCCATGGAC 3′) with 2 mM $MgCl_2$. This primer pair amplified 1,313 bp containing the RT/RNAseH domains frequently used in *Badnavirus* phylogeny.

The following thermal cycling profiles were used: *trnK-matK* region: 5 min of denaturation at 94 °C, 30 cycles of (30 sec of denaturation at 94 °C, 30 sec annealing at 58 °C, 1 min 15 sec of extension at 72 °C) and 10 min for final extension; the *trnL-trnF* region: 2 min of denaturation at 94 °C, 30 cycles of (30 sec of denaturation at 94 °C, 30 sec annealing at 52 °C, 3 min of extension at 72 °C) and 10 min for final extension; PhyloEPRV-Gf: 5 min of denaturation at 95 °C, 35 cycles of (30 sec of denaturation at 95 °C, 30 sec annealing at 50 °C, 1 min 30 sec of extension at 72 °C) and 10 min at 72 °C for final extension. See below for detailed PCR mixture. Sequencing was performed by Cogenics Genome Express SA (Grenoble, France) in reverse for *trnL-trnF* and in both strands for PhyloEPRV-Gf, *matK*. Genbank accession numbers available upon acceptance of the manuscript (will appear as supplementary data).

Phylogenetic inferences

Sequences were aligned using ClustalW (Thompson et al. 1994) implemented in Bioedit (Hall 1999) and corrected manually when necessary. The primer sequences were then discarded from the alignments. The software DAMBE version 4.5.20 (Xia and Xie 2001) was used to detect substitution saturation in each of the six alignments following a previously described method (Xia et al. 2003). For this purpose, the percentage of invariant sites was first estimated by PhyML v2.4.4 software (Guindon et al. 2005) using a GTR + I substitution model with 8 categories of Gamma parameter. Expected saturation index was given for symmetric tree topology.

Maximum likelihood phylogenies were inferred using PHYML v2.4.4 software. A GTR model with 8 categories of Gamma parameter and with a fixed proportion of invariable sites (0.01%) was used. 500 bootstrap iterations were performed to assess the robustness of trees topologies.

Trees topologies were also computed with MrBayes 3.1.2 (Huelsenbeck and Ronquist 2001) by running 5 chains and 10^6 generations using the default priors of the GTR model. Bayesian posterior probabilities were calculated from majority-rule consensus of trees sampled every 20 generations once the Markov chains had become stationary (determined by empirical checking of likelihood values).

Sequences of the RT/RNAseH region of ORF3 of BSVs species BSImV and BSGFV were retrieved from public databank. Episomal BSImV virus (BSImVUg) clones 10.1 (emb: AJ968444), 20.3 (emb: AJ968445), 21.6 (emb: AJ968447), 39.1 (emb: AJ968449), 49.2 (emb: AJ968450) and 49.6 (emb: AJ968451) and BSGFV virus clones GFUg50.1 (emb: AJ968435), 54.4 (emb: AJ968439), 54.6 (emb: AJ968441), 54.8 (emb: AJ968442), 54.2 (emb: AJ968438), 54.5 (emb: AJ968440) originated from BSV epidemics in Uganda (Harper et al. 2005). BSGFV episomal clones Col20 (gb: EU076416), Col23 (gb: EU076417), Col28 (gb: EU076418), Col30 (gb: EU076420), Col31 (gb: EU076421), Col32 (gb: EU076422), and Col34 (gb: EU076423) originated from Colombian epidemics. BSImV EPRVs clones Bat27 (gb: AY189426) and KT32 (gb: AY452264) were amplified from healthy *Musa balbisiana* cv. Pisang Batu (genotype BB) and from banana cv. 'Klue Tiparot' (genotype ABB), respectively (Geering et al. 2005). Complete BSV genomic sequences of BSAcVNV (gb: AY750155) (Lheureux et al.

2007), of BSGFV (gb: AY493509) (Gayral et al. 2008) and BSImV (Gayral et al; unpublished) were also used in this study.

Microsatellite genotyping

The 19 microsatellites loci used in this study were independent (I. Hippolyte, unpublished data). They were developed in *M. acuminata* cv. 'Gobusik' (Kaemmer et al. 1997; Lagoda et al. 1998) and *M. balbisiana* cv. 'Pisang Klutuk Wulung' (I. Hippolyte, unpublished data) (Table 2). Genotyping was performed using fluorescently-labelled polymerase chain reaction (PCR). PCR were carried out in 25 µl of a mixture containing approx. 25 ng of DNA, 0,14 µM reverse primer and 0.12 µM M13-tailed forward primer, 1.25 U of GoTaq Flexi DNA Polymerase (Promega, Madison, WI, USA), 0.1 mM each dNTP, 1.5 mM MgCl2, 5 µl of 5 x Colorless GoTaq flexi buffer and 0,16 µM of M13 primer fluorescently labeled with hexachlorocarboxyfluorescein (HEX) and carboxyfluorescein (6-FAM) (Eurogentec, Maastricht, the Netherlands). An initial denaturing step of 2 min at 94 °C was followed by 40 cycles (94 °C for 30 s, 53 °C for 30 s and 72 °C for 1min) and 10 min at 72 °C. PCR products were diluted (1/5; depending on their concentration). Sizes of amplified fragments were measured by a MegaBACE capillary sequencing machine (Amersham Biosciences, Freiburg, Germany). Alleles were scored using MegaBACE Genetic Profiler v1.0 software (Amersham Biosciences). Alleles included in final consensus genotypes were observed at least twice. Two samples with known genotypes served as positive controls and were included in each run of 96 PCRs to standardize genotyping across experiments.

Table 2: Description of the microsatellite loci used and summary of the allelic variation

Locus ID	GenBank Acc. Nb.	Repeat motif	No. of alleles			Alleles size range	He[e]		Ho[f]
			M. balbisiana	other *Musa*.	Total		*M. balbisiana*	other *Musa*	
mMaCIR08[c]	X87264	$(TC)_6N_{24}(TC)_7$	5	12	14	251-287	0.573	0.962	0.623
mMaCIR307[d]	AM950533	$(CA)_6$	1	4	4	162-170	0.000	0.723	0.083
mMaCIR07[c]	X87258	$(GA)_{13}$	6	12	14	148-188	0.470	0.939	0.500
mMaCIR13[c]	X90745	$(GA)_{16}N_{76}(GA)_8$	6	11	14	266-307	0.676	0.945	0.564
Ma3-90[b]	NA[a]	NA[a]	4	11	14	142-177	0.554	0.932	0.444
mMaCIR01[c]	X87262	$(GA)_{20}$ $(CA)_5GATA$	6	10	14	248-329	0.609	0.932	0.586
mMaCIR39[c]	Z85970	$(GA)_5$	6	10	13	295-388	0.681	0.928	0.653
mMaCIR260[d]	AM950515	$(TG)_8$	3	6	7	208-256	0.608	0.848	0.358
mMaCIR03[c]	X87263	$(GA)_{10}$	2	7	7	116-148	0.516	0.818	0.367
mMaCIR40[c]	Z85977	$(GA)_{13}$	5	11	15	170-232	0.750	0.955	0.530
mMaCIR150[d]	AM950440	$(CA)_{10}$	3	7	8	253-271	0.495	0.856	0.425
mMaCIR214[d]	AM950480	$(AC)_7$	3	3	4	122-130	0.611	0.572	0.458
mMaCIR264[d]	AM950519	$(CT)_{17}$	1	9	9	240-280	0.000	0.932	0.167
mMaCIR152[d]	AM950442	$(CTT)_{18}$	6	7	11	160-196	0.564	0.873	0.427
mMaCIR164[d]	AM950454	$(AC)_{14}$	4	10	13	300-436	0.716	0.972	0.286
mMaCIR196[d]	AM950462	$(TA)_4(TC)_{17}$ $(TC)_3$	4	10	11	162-198	0.737	0.945	0.535
mMaCIR231[d]	AM950497	$(TC)_{10}$ $(TA)_4CA$	2	10	12	240-280	0.523	0.902	0.344
mMaCIR45[c]	Z85968	$(CTCGA)_4$	2	9	9	278-296	0.329	0.939	0.275
mMaCIR24[c]	Z85972	$(TC)_7$	4	9	9	239-289	0.655	0.809	0.432

[a]NA: Not availiable
[b]Kaemmer et al., 1997 (Kaemmer et al. 1997)
[c]Lagoda et al., 1998 (Lagoda et al. 1998)
[d]Hippolyte et al., in prep.
[e]He is the gene diversity as computed using FSTAT
[f]H0 is the observed heterozygosity

Genetic distances and microsatellite genetic variation

We used the inter-individual microsatellite genetic distance 'Simple Matching' implemented in DARwin v5.0.155 software (Perrier, X., Jacquemoud-Collet, J.P. (2006); http://darwin.cirad.fr/darwin). This distance, also called '1- Dps' distance, takes into account the proportion of shared alleles. Experimental data suggested that the proportion of shared alleles was effective in obtaining correct genealogical relationships (Harr et al. 1998; Muir et al. 2000). 1,000 bootstrap replicates were performed with DARwin softwares. Distance matrices were then used to construct dendrograms with the NJ algorithm implemented in DARwin software. The program FSTAT v.2.9.3.2 (Goudet 1995) (http://www2.unil.ch/popgen/softwares/fstat.htm) was used to compute values of standard genetic diversity indices.

PCR screen of EPRV

Distribution of BSGFV and BSImV EPRVs among the genus *Musa* was performed by amplifying the 8 and the 5 characterized junctions of BSGFV EPRV and BSImV EPRV in cv. PKW, respectively (Table 3).

PCR-based genotyping errors resulting in no amplified PCR product were first lowered by two complementary approaches. We first detected low DNA quantity or problems in extraction quality by performing in parallel a PCR amplification of the housekeeping *actin* gene in all samples [Actine1F (5'-TCCTTTCGCTCTATGCCAGT-3') and Actine1R (5'-GCCCATCGGGAAGTTCATAG-3') at Tm = 58°C for 25 cycles and with 1.5 mM $MgCl_2$, see above] (data not shown). PCRs or DNA extraction was repeated if no acceptable amplification of *actin* gene was observed. Then, we verified that the absence of amplification was not solely explained by a SNP in primer sequence if a PCR marker still would not amplify. To do so, we performed a second PCR that amplified the same fragment but with slightly more external primers (labeled 'BIS' in Table 3). Finally, genotyping errors due to unspecific PCR amplification were also reduced. We first ensured that episomal BSImV and BSGFV viruses were not detected by any of the EPRV primers, which would have also resulted in false positives. Then, the few PCR products showing unexpected product sizes were systematically cloned into TOPO-TA (Invitrogen, Carlsbad, CA) according to the manufacturer's instructions, sequenced and then discarded if they were not BSV-related sequences (data not shown).

Table 3: primers used in the PCR screen of BSGfV and BSImV integrations

EPRV amplified	Primer name	Primer sequence (5'-3')[a]	Expected product size (base-pairs)
BSGFV[b]	VM1-F	TTGTCCAAAATCTGCTCGTG	481
	VM1-R	TGTAATTCCTGCTCCTGCAA	
	VM2-F	TTCTCCCTTTTCGATCCGTA	374
	VM2-R	TTTTGATGCATCTCCAGCAG	
	VM2bis-F	GAGGCCCTTATGCATTGTTG	159
	VM2bis-R	TCGACCGTACCGATATCCTC	
	VV1-F	ACAGCTCCAGGAGATTGGAA	268
	VV1-R	CTGAAGTGTGCCTGTGGAGA	
	VV2-F	TCTGAGATCTCCAGCCAGGT	639
	VV2-R	GACAGTTCCAGCACAGCAGA	
	VV2bis-F	GCTGGCAGTGGAATTCAGTT	395
	VV2bis-R	CATGGTGGGAGAAGAGGAAG	
	VV3-F	TTGCCAAGAATTCCTCCAAG	376
	VV3-R	AAGTTCTTGTCGGCAAGGTG	
	VV4-F	GAGCAACACGAGTCAACGAA	784
	VV4-R	TCTCCACAGGCACACTTCAG	
	VV4bis-F	GGAAAACTCTGGGTTGGTGA	766
	VV4bis-R	GGAGTACGGCATTTTCTCCA	
	VV5-F	CCATGGAGGTTGACCTGTCT	628
	VV5-R	ACCCCTCTGTCTTCCCAACT	
	VV5bis-F	CGCACCTTCATCACAGAAGA	588
	VV5bis-R	TACCAGATGGGGAGAAATCG	
	VV6-F	GCATGAAGCATGACTGGAGA	264
	VV6-R	AATGCATAAGGGCCTCGAAT	
	VV6bis-F	AGGCCACTACGCATCAGAAT	712
	VV6bis-R	GGCCTCGAATTTATCATTGG	
BSImV[c]	Musa/F2-F	ACTCAGCAAAGGCAAGCAGT	561
	Musa/F2-R	TCTGGTGTGAGTTTTAATAATACCG	
	Musa/F2bis-F	AGCTGAAGTGATGCGAACCT	937
	Musa/F2-R	-	
	F1/F3-F	TTCGGTATTATTAAAACTCACACCA	490
	F1/F3-R	GCTGCTAACTGAGGATAATCGAA	
	F1/F3-F	-	630
	F1/F3bis-R	TTCTTGGGGTACTGGTTTCG	
	F3/F4-F	TCCCACGCAAGCTTACTTCT	600
	F3/F4-R	GAAGCTGTCCAAGCCTATATCA	
	F3/F4bis-F	GGTGCAAATCAGAGTCATGC	987
	F3/F4-R	-	
	F4/F5-F	TGGACAGCTTCTGGTGTGAG	540
	F4/F5-R	AGCAGCTACAACCCTGGAGA	
	F4/F5-F	-	927
	F4/F5bis-R	AGCATCCGCTTTGGAGACTA	
	F5/Musa-F	GTATGGTTCTTGCCCGATGA	594
	F5/Musa-R	TCGTGCAGACCCCTTACTCT	
	F5/Musabis-F	CCACCTGGTATCCCTGAAGA	905
	F5/Musabis-R	TGTCAAGCTGTTGGTTGCTC	

[a]Primer annealing temperature : 60 °C

[b]Primer pairs 'VM' amplify the junction between *Musa* genome and EPRV, primer pairs 'VV' amplify internal fragment junctions within BSGFV ERPV

[c]PCR F*i*/F*j* amplify the junction between fragment *i* and fragment *j* of BSImV EPRV

PCRs were carried out with 5-20 ng of DNA, 20 mM Tris–HCl (pH 8.4), 50 mM KCl, 100 mM of each dNTP, 1.5 $MgCl_2$, 10 pmol of each primer and 1U Taq DNA polymerase (Eurogentech, Seraing, Belgium) in 25µl. PCR products were visualized under UV light after migration of 10 µl of PCR products on a 1.5 % agarose gel in 0.5 x TBE (45 mM Tris-borate, 1 mM EDTA, pH = 8) stained with ethidium bromide.

Southern blot hybridization

Ten µg of genomic DNA of *Musa* sp. was digested overnight at 37 °C in 100 µl with 20 units of EcoRI (Promega, Madison, WI, USA), 10 µl of 10 x buffer (Promega, Madison, WI, USA), 1 µl of 100 x BSA (Promega, Madison, WI, USA). Following concentration in a 20 µl volume by centrifugal evaporation in a Speedvac, digested DNA was separated during 6 h at 80 V on a 0.8 % agarose gel in 0.5 x Tris–borate–EDTA (TBE) buffer. The gel was soaked in 0.25 N HCl for 15 min, then in 0,4 N NaOH for 15 min. Following ethidium bromide staining and UV light visualization, the DNA was transferred to a positively charged nylon membrane (Hybond N^+; Amersham Biosciences, Arlington Heights, IL) by upward capillary transfer and immobilized by UV irradiation.

Two BSGFV clones [pCR-TOPO-1.2 (1,262 bp) and pCR-TOPO-6 (6,001 bp)] covering the entire BSGFV genome were used as probes in equimolar quantity. Probes were linearized with EcorI, and BamHI + EcoRV, respectively, and gel-purified with Wizard^R SV Gel and PCR Clean Up System (Promega, Madison, WI, USA). The excised DNA was then labeled during 1 h at room temperature with [α-^{32}P] dCTP using a Prime-a-gene Labeling System (Promega, Madison, WI, USA) following manufacturer's instructions.

Membranes were pre-hybridized overnight at 65°C in a solution of 5 x SSC containing 2.5X Denhardt's reagent, 50 mM Tris-HCl pH 8, 10 mM EDTA, 0.2% SDS and 100µg/ml of 5 min boiled herring sperm DNA. Hybridization with heat denatured probe was done during 4 h at 65°C in the same solution as pre-hybridization with 10 x Dextran sulfate.

Membranes were then washed twice for 10 min with 0,5 x SSC, 0.1 % SDS at 65°C, followed by three 30 min washes with 0.2 x SSC and 0.1 % SDS at 65°C. Bound

probes were detected using a PosphorImager system (Storm). Membranes were stripped of probe by washing in boiling 0.1% SDS three times, then rinsing in 2 x SSC.

Results

Morphotaxomonic analyses divided *Musa* genus into four sections: *Eumusa*, *Rhodochlamys*, *Australimusa* and *Callimusa*. Our sample of 32 accessions emphasized on *Eumusa* section where *Musa balbisiana* belonged (Table 1).

Chloroplast molecular phylogeny of *Musa*

Two accessions were added to the sample of 32 accessions: the 'Grande Naine' reference, a cultivar of the Cavendish subgroup mainly used in commercial plantation and *Ensete glaucum* as outgroup.

We first tested if our alignments were free of substitution saturation, i.e. suitable for phylogenies. It is assumed that phylogenetic information is essentially lost when the observed saturation index is equal or larger than half of full substitution saturation (Xia 1999). We estimated the expected saturation indices assuming half of the full substitution saturation and compared it to the observed saturation indices. Substitution saturation is detected when the observed indices are higher than the expected indices. No saturation was detected in any of the three alignments used in this study: the concatenation of *matK* gene and *TrnL-F* region, and the RT/RNAseH region of BSGFV and BSImV ($p < 10^{-4}$) (Table 4).

Table 4: Detection of saturation substitution with the program DAMBE

Alignments	Observed Saturation index (Iss)	Expected Saturation index (Iss.cSym)	*T*	DF	*p*-value[a]
Concat-*matK_TrnL-F*	0.1297	0.8011	35.210	2158	0.0000
BSGVF	0.1264	0.7167	37.959	389	0.0000
BSImV	0.0865	0.7342	59.160	459	0.0000

[a] The statistical significance of the difference between observed and expected saturation indexes was assessed with a two-tailed test

Figure 1 shows the ML chloroplast phylogeny inferred from a combined alignment of 2159 nt made from the concatenation of *matK* and *trnL-trnF* alignments. When

used separately, *matK* phylogeny and *trnL-trnF* phylogeny were congruent with the combined analysis (data not shown). Bayesian analysis using the combined alignment produced a tree with very similar topology and branch length than using ML (data not shown).

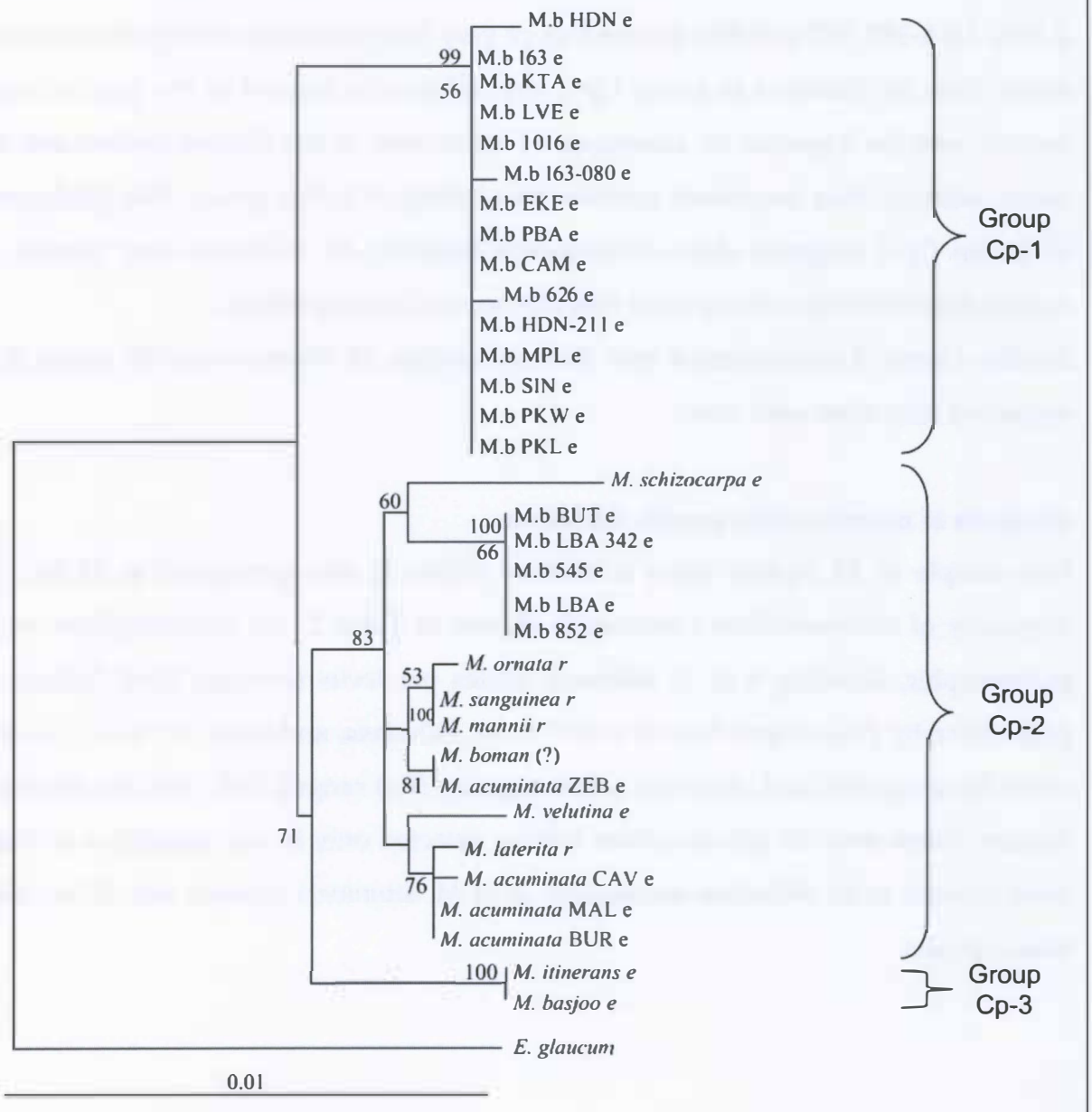

Figure 1: Chloroplast ML phylogeny of *Musa* based on *matk* gene and *trnL-trnF* region concatenated in a single alignment of 2.1 kbp. Bootstrap values over 50 % (percentage from 500 replicates) are shown at the left of the nodes. Lowercase letters after species or accession names indicate the *Eumusa* (e) or *Rhodochlamys* (r) section, respectively. *Ensete glaucum* sequence was used as outgroup.

The phylogeny separated the accessions into three well supported groups noted Cp-1 to Cp-3 (Cp for chloroplast). Group Cp-1 contains 15 out of the 20 accessions of *Musa balbisiana* species. The intra-group diversity was very low as attested by the short branches. Only three accessions, HDN, I63-080 and 626 displayed a low polymorphism (99.8 to 99.9 % nucleotide identity).

The 5 remaining accessions of *M. balbisiana* species (acc. 'Butuhan' (BUT), 'Los Baños' (LBA), LBA-342, 545 and 852) showed no or very few nucleotide divergence between them. They all clustered in group Cp-2 with all species studied of the *Rhodochlamys* section, and the 2 species *M. acuminata*. *M. schizocarpa* of the *Eumusa* section and *M. boman* with no clear taxonomic position also belonged to this group. The phylogeny of group Cp-2 suggests close relationships between *M. acuminata* and species of section *Rhodochlamys* and supports that this section is paraphyletic.

Finally, Group 3 encompassed two *Eumusa* species: *M. itinerans* and *M. basjoo* that appeared very close each other.

Analysis of microsatellite genetic variation

Our sample of 32 diploid *Musa* accessions (Table 1) was genotyped at 19 loci. A summary of microsatellites variation is shown in Table 2. All microsatellites were polymorphic, showing 4 to 15 different alleles per locus (average 10.6). Values of gene diversity (H_e) ranged from 0 to 0.75 in *M. balbisiana*, and from 0.57 to 0.97 in the other *Musa* species, and observed heterozygosity (H_o) ranged 0.08 - 0.65 for the total dataset. There were 59 private alleles (alleles detected only in one sample), 4 of them were present in *M. balbisiana* accessions, 18 in *M. acuminata* species, and 37 in other *Musa* species.

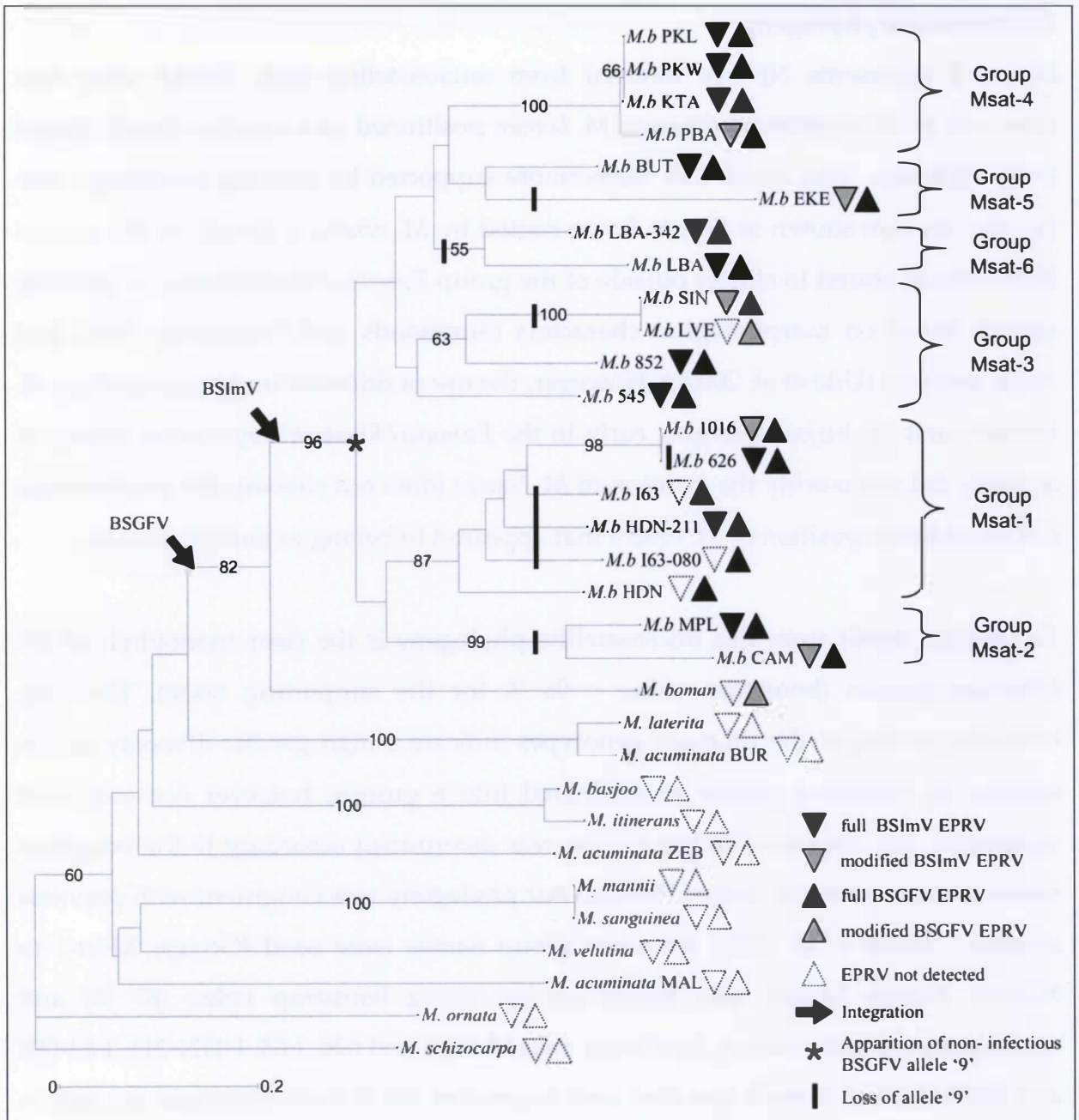

Figure 2: Neighbor-joining tree reconstructed from Simple Matching genetic distance from 19 microsatellite loci. Bootstrap values over 50 % (percentage from 1000 replicates) are shown at the left of the nodes. Distribution of EPRVs was investigated with 8 PCR markers for BSGFV EPRV (upward triangles) and 5 PCR markers for BSImV (downward triangles). Full-length ERPV are represented with black triangles, modified EPRV with empty triangles and absence of EPRV with dotted triangles (see Table 1 for details). Upward and downward arrows indicate inferred BSGFV and BSImV integration events, respectively. Presumed apparition of the non-infectious BSGFV EPRV allele '9' and its loss in *Musa* genomes are represented by asterisk and vertical lines, respectively. Tree was rooted with *Musa ornata* sequence.

Microsatellite phylogeny

Figure 2 represents NJ tree inferred from microsatellite data. Unlike what was observed in chloroplast phylogeny, *M. boman* positioned as a species closely-related to *M. balbisiana*. This result was furthermore supported by a strong bootstrap value (82 %). The tree shown in Figure 2 was rooted by *M. ornata*, a species of the section *Rhodochlamys* found to cluster outside of the group *Eumusa/Rhodochlamys* in phenetic studies based on morphological characters (Simmonds and Weatherup 1990) and AFLP markers (Ude et al. 2002a). However, the use of different outgroups such as *M. itinerans* and *M. basjoo* diverging early in the *Eumusa/Rhodochlamys* group (Wong et al. 2002) did not modify the position of *M. boman* (data not shown). We are therefore confident in the position of *M. boman* that appeared to belong to *Eumusa* section.

The second result from this microsatellite phylogeny is the clear monophyly of *M. balbisiana* species (bootstrap value = 96 % for the supporting node). The long branches leading to the different genotypes indicate a high genetic diversity of this species. *M. balbisiana* species is structured into 6 groups, however not very well supported. The definition of the groups was determined according to the bootstrap value of the supporting nodes. Because our phylogeny was congruent with previous studies (Carreel et al. 2002) the same group names were used (Groups Msat-1 to Msat-6). Group Msat-1 was supported by strong bootstrap value (87 %) and encompassed six accessions: the closely related 1016 and 626, I 63, HDN-211, I 63-080 and HDN. Group Msat-2 was also well supported (99 %) and contained accessions CAM and MPL. The last 4 groups were less supported. Group Msat-3 contained the closely-related SIN and LVE accessions, as well as accessions 852 and 545. Group Msat-4 contained PBA accession and the closely-related KTA, PKL and PKW accessions. Group Msat-5 encompassed the two BUT and EKE accessions. Finally, group Msat-6 contained the two LBA and LBA-342 accessions.

Distribution of BSImV and BSGFV integrants among *Musa* genus

To study the distribution of BSImV and BSGFV EPRV among the *Musa* genus, we first had to develop new tools that enabled to detect specifically these EPRVs. The structure and biological properties of BSGFV EPRVs from *M. balbisiana* cv. PKW were

described in Gayral et al. (2008). Similarly, EPRV BSImV was recently sequenced from cv. PKW and its description will be presented elsewhere (Gayral et al. unpublished data). BSGFV and BSImV EPRVs are independent integrations of two distinct viral species, but they nevertheless show a similar organization in *M. balbisiana* cv. PKW. These EPRVs, ranging from 13.3 to 15.8 kbp in size, are composed of juxtaposed viral fragments in both orientations. To genotype these EPRVs, we designed 8 PCR markers for BSGFV and 5 PCR markers for BSImV that amplify junctions of the rearranged fragments. Therefore, each set of markers produced a unique signature of BSImV and BSGFV EPRV as known in cv. PKW. Furthermore, any of these PCRs cross-amplified with the episomal form of the corresponding viruses.

For BSGFV ERPV, 'VV' primers (junction virus-virus) amplify the 5 internal junctions of the 6 fragments present in the infectious allele (EPRV-7). Primer pair VV5 amplifies the junction involved in the supplementary fragment of the non infectious allele (EPRV-9) found in *M. balbisiana* cv. PKW. Two additional 'VM' primers (junction virus-*Musa*) amplify the 5' and 3' boundaries of EPRV-9 (See Table 3). The same approach was used concerning BSImV integrant in cv. PKW. BSImV EPRV is homozygous and is composed of 5 fragments in both orientations (Gayral al, unpublished data). As several fragments were very short (30 and 27 nucleotides for fragment I and II, respectively), it was not possible to amplify each junction separately. Three PCR markers (primers F1/F3, F3F4 and F4/F5; see Table 3) specific to the internal fragments of this EPRV were finally designed and used in this study. Two additional PCR markers (Musa/F2 and F5/Musa) amplifying 5' and 3' boundaries of BSImV EPRV, were specific of the integration locus in *Musa* genome.

Details of the distribution of EPRVs markers after genotyping are shown in Table 1, and results are summarized in Figure 2 regarding to the host microsatellite phylogeny. The presence of a full-length EPRV in distantly-related accessions is interpreted as a shared and derived (synapomorphic) character state. Indeed, two independent integrations would not result in the same fragment organization of EPRVs because they are assumed to originate randomly, and would be distinguished

by absence of PCR markers coupled with amplification at unexpected size, which was never observed. Then, modified EPRVs, as represented in Figure 2, were revealed by one or several missing markers. Because modified EPRVs originated from independent mutations in distinct EPRV fragments, they are derived but not shared character states, and are therefore not symplesiomorphic.

BSImV EPRV was restricted to the species *M. balbisiana* only, and was not detected in any of other species. The BSImV integration event probably occurred after the speciation of *M. boman* and *M. balbisiana* species, and before the diversification of *M. balbisiana* since all the 6 groups carried this integration (Figure 2). An important polymorphism of EPRV structure was observed among *M. balbisiana* genotypes. Although 11 accessions carried the full-length EPRV as described in cv. PKW, 5 accessions presented a variable number of missing PCR markers at different positions within EPRV (Table 1), attesting a modification or truncation of the integrations. Modified EPRVs were observed in 5 groups (groups Msat-1 to Msat-5), indicating that degradation of BSImV EPRV would have occurred continuously during evolution of *M. balbisiana*. Finally, none BSImV EPRV markers were amplified in the genomes of 4 accessions, which probably lack this EPRV. So, two independent loss of BSImV EPRV would be necessary to explain the observed distribution and the absence of BSImV EPRV in some *M. balbisiana* genotypes. A first loss would have occurred in Group Msat-3 (accession LVE). A second loss would have occurred in Group Msat-1 in a putative common ancestor of accessions HDN, I 63 and I 63-080. These accessions did not form a clade in Figure 2, however since the nodes are poorly robust, one can hypothesize they form a monophyletic group excluding accession HDN-211. These results suggest an unstable structure of BSImV EPRV.

Unlike BSImV, BSGFV EPRV was not restricted to the B genome (*M. balbisiana*). One other *Eumusa* species, *M. boman*, also carried an altered version of this EPRV (Figure 2). Half of the markers specific for internal rearrangements were present in *M. boman*, as well as the two markers specific for the locus of integration into the *Musa* genome (Table 1), indicating that this EPRV derived from the same EPRV BSGFV present in *M. balbisiana* cv. PKW. To verify that BSGFV sequence is integrated into the genome of *M. boman*, and is absent from other *Eumusa* and *Rhodochlamys* species, southern

blots were performed on EcoRI-digested total genomic DNA of 8 accessions hybridized with two probes covering the entire BSGFV genome. Hybridization was observed with DNA from *M. boman* but its fingerprint pattern was different from those obtained with *M. balbisiana* cv. PKW DNA. Two bands out of 5 were missing (Figure 3). Hybridization was not observed with DNA neither from *Musa itinerans*, *M. basjoo*, *M. laterita*, *M. schizocarpa*, nor from the two *M. acuminata* subsp. *malaccensis* and *burmanicoides* analyzed. BSGFV EPRV was detected in all *M. balbisiana* accessions and was much less polymorphic than BSImV EPRV. Only one accession (LVE) lacks one internal PCR marker. This loss may be very recent since the sister taxa SIN still carried this marker.

The only polymorphism observed in BSGFV EPRVs concerned the difference between the two alleles (EPRV-7 and EPRV-9) present in cv. PKW in which the allele EPRV-7 is infectious whereas the allele EPRV-9 is not. Comparison between both alleles suggested that allele EPRV-9 derived from the allele EPRV-7 by accumulation of deleterious mutations in ORFs, and to an internal tandem duplication of 2.3 kbp (Gayral et al., 2008). Whereas we expected EPRV-9 would be advantageous for the host since it is not associated with the release of BSV particles, we did not observe a dynamic of loss or fixation during the diversification of *M. balbisiana* species. For instance, bananas of Groups Msat-2 and Msat-6 carried the infectious allele only and those of the other groups carried either the infectious allele only, or both alleles (Table 1). This result summarized on Figure 2 suggests that EPRV-9 allele appeared before the divergence of the 6 groups, and that independent loss of this allele occurred at least 6 times during *M. balbisiana* evolution.

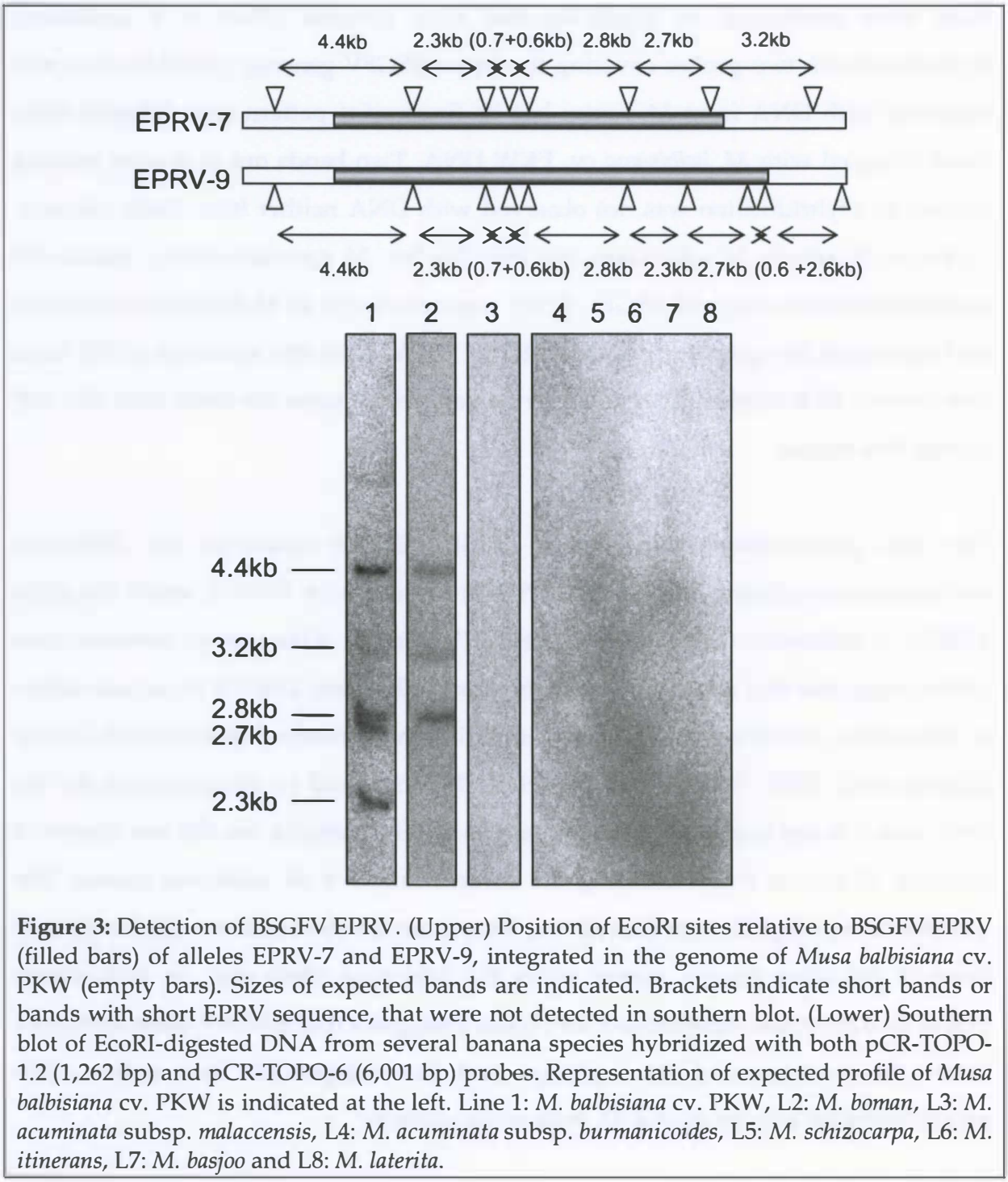

Figure 3: Detection of BSGFV EPRV. (Upper) Position of EcoRI sites relative to BSGFV EPRV (filled bars) of alleles EPRV-7 and EPRV-9, integrated in the genome of *Musa balbisiana* cv. PKW (empty bars). Sizes of expected bands are indicated. Brackets indicate short bands or bands with short EPRV sequence, that were not detected in southern blot. (Lower) Southern blot of EcoRI-digested DNA from several banana species hybridized with both pCR-TOPO-1.2 (1,262 bp) and pCR-TOPO-6 (6,001 bp) probes. Representation of expected profile of *Musa balbisiana* cv. PKW is indicated at the left. Line 1: *M. balbisiana* cv. PKW, L2: *M. boman*, L3: *M. acuminata* subsp. *malaccensis*, L4: *M. acuminata* subsp. *burmanicoides*, L5: *M. schizocarpa*, L6: *M. itinerans*, L7: *M. basjoo* and L8: *M. laterita*.

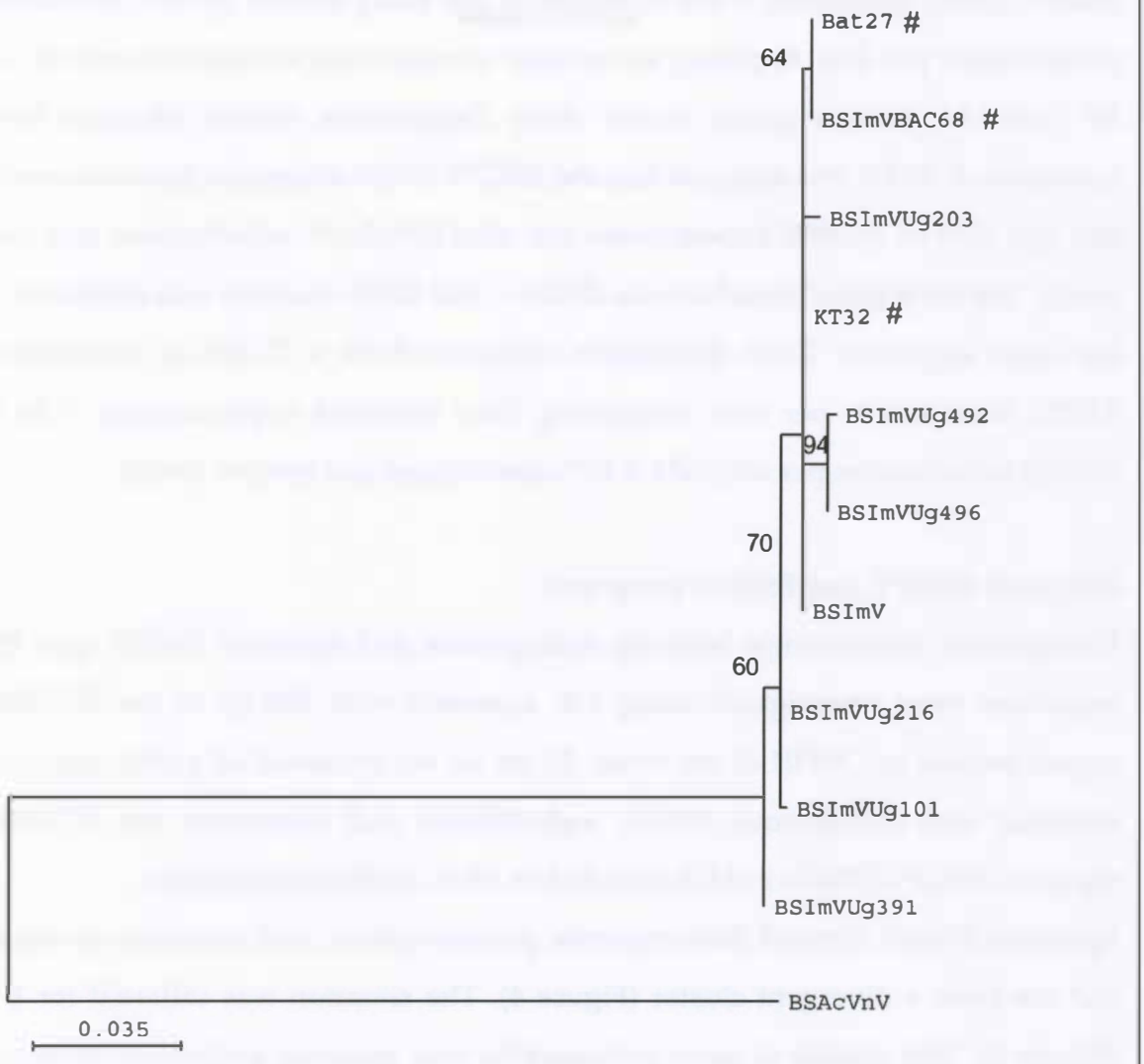

Figure 4: Maximum Likelihood phylogeny of RT/RNaseH region (550 bp) of endogenous and episomal sequences of BSImV. Hash signs indicate EPRV sequences, other are episomal sequences. EPRV sequence present in *M. balbisiana* cv. PKW is BSImVBAC68 (Gayral et al, unpublished data). Bootstrap values over 50 % (percentage from 500 replicates) are shown at the left of the nodes. The tree is rooted with the closely-related species BSAcVNV (episomal BSV species).

Age of integration and allelic divergence of BSGFV EPRV

In a previous study, we showed that EPRV BSGFV was imbedded inside a Ty3/Gypsy retroelement. One hypothesis to explain this situation was a simultaneous integration by retrotransposition of a retroelement/BSGFV chimera formed after recombination at the RNA level (Gayral et al. 2008). The divergence between the two long terminal repeats sequences (LTRs) of this retroelement was estimated with maximum likelihood using the HKY model implemented in PhyloWin (Galtier et al. 1996) and was used to determine the date of integration. For

BSGFV EPRV, estimation of the substitution rate along the 351 bp of LTRs was 0.0058 substitutions per site. Applying an average synonymous substitution rate of 4.5 per 10^9 years for nuclear genes in the order Zingiberales (where *Musaceae* belongs) (Lescot et al. 2008), this suggests that the BSGFV EPRV integrated approximately 0.64 My ago (MYA) (0.0058 substitutions per site/(2*4,5.10^{-9} substitutions per site per year)). The divergence time between EPRV-7 and EPRV-9 alleles was estimated using the same approach. Their divergence estimated from a 13,280 bp alignment was 0.0022 substitution per site, suggesting they diverged approximately 0.24 MYA (0.0022 substitutions per site/(2*4,5.10^{-9} substitutions per site per year)).

Origin of BSGFV and BSImV integrants

Phylogenetic relationships between endogenous and episomal BSGFV and BSImV sequences were investigated using ML approach with 550 bp of the RT/RNaseH region present in ORFIII of the virus. To do so, we retrieved all public sequences of episomal and endogenous BSImV and BSGFV, and sequenced the RT/RNaseH region of BSGFV EPRVs in *M. boman* and in 20 *M. balbisiana* accessions.

Episomal BSImV showed little sequence polymorphism, and endogenous sequences did not form a divergent cluster (Figure 4). The situation was different for BSGFV (Figure 5). This species is more polymorphic and episomal sequences formed three distinct groups (named GF-1 to GF-3). Our results showed that EPRV derived from Group GF-1 viruses. In this group, the episomal sequence of 'BSGFV' and EPRVs emerged from the same node, suggesting a recent common ancestor. Furthermore, the branches leading to EPRVs were particularly short (< 0.002 substitution/site on average), whereas those of the episomal sequence (BSGFV) were slightly longer (0.006 substitution/site). This result indicated a slower rate of evolution for EPRV compared to episomal sequences.

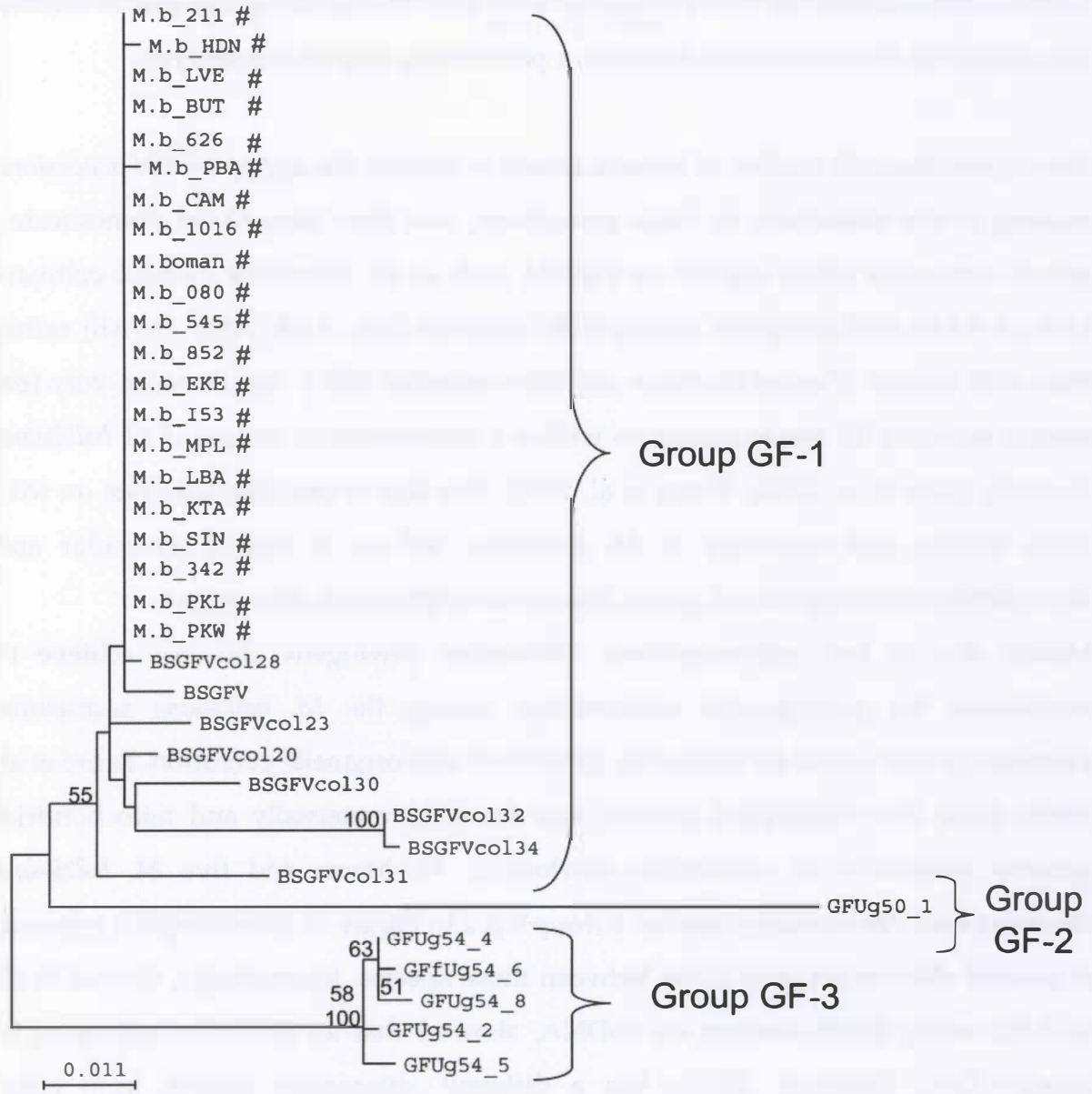

Figure 5: Maximum Likelihood phylogeny of RT/RNaseH region (550 bp) of endogenous and episomal sequences of BSGFV. Hash signs indicate EPRV sequences, other are episomal sequences. Phylogenetic groups among this viral species are indicated with braces. Bootstrap values over 50 % (percentage from 500 replicates) are shown at the left of the nodes.

Discussion

The aim of this study was to investigate the origin and the evolution of two infectious BSV EPRVs in the *Musa* genus. To do so, we followed the distribution of the infectious integrations of two *Banana streak virus* species BSImV and BSGFV in a large sample of 61 accessions using 13 specific PCR markers. By focusing on 32 accessions including 20 *M. balbisiana* and several closely-related banana species, we were able to dissect their early stages of evolution in relation to the phylogeny of the

hosts. Reconstructing the host phylogeny from two chloroplast genes and 19 nuclear microsatellites loci constituted therefore a preliminary step of our analysis.

Recent phylogenetic studies of banana aimed to classify the approx. 2,000 accessions existing in the collections of *Musa* germplasm, and thus focused on domesticated genetic resources either diploid or triploid such as *M. acuminata* derived cultivars (AA or AAA) and polyploid interspecific cultivars (AB, AAB, ABB, AAAB) rather than wild species (Heslop-Harrison and Schwarzacher 2007). Additionally, very few studies included *M. boman* species as well as a representative sample of *M. balbisiana* diversity (Ude et al. 2002b; Wong et al. 2002). The lack of phylogenetic data on wild *Musa* species and especially of *M. balbisiana*, led us to realize molecular and microsatellites phylogenies of genus *Musa* with emphasis on this species.
Mainly due to low polymorphism, chloroplast phylogeny did not achieve to reconstruct the phylogenetic relationships among the *M. balbisiana* accessions. However, it was useful for observing gene flows and organelle evolution. Fauré et al. (1994) show that chloroplast genome was inherited maternally and mitochondrial genome paternally. In chloroplast phylogeny, *M. boman* and five *M. balbisiana* clustered near *Rhodochlamys* species (Group Cp-2 in Figure 1), indicating the presence of ancient chloroplast gene flows between those species. Interestingly, Carreel et al. in 2002, using RFLP markers on cpDNA, showed that an accession belonging to groups Cp-2 'Butuhan' (BUT), has a different chloroplast pattern from other 'conventional' *M. balbisiana*, shared with the *Australimusa* species. This banana was suspected to be a back-cross of *M. balbisiana* as a father on a *M. textilis* x *M. balbisiana* hybrid, therefore explaining its intermediary position found in chloroplast phylogeny in this study. One must however moderate the presence of extended chloroplast gene flow since several important nodes of group Cp-2 are not supported.

The repertoire of microsatellite analysis has recently been successfully extended beyond population genetics and was successfully used to infer accurate phylogenetic relationships (Schlotterer 2001). We conducted our analysis using 19 microsatellite loci. Several recent studies using approximately the same number or fewer loci

provided strong phylogenetic signals and accurate phylogenetic reconstruction of the mice *Peromyscus maniculatus* (Chirhart et al. 2005) (12 loci), of the tribe *bovini* (Ritz et al. 2000) (20 loci), of oaks species (Muir et al. 2000) (20 loci) or of lizard populations (Richard and Thorpe 2001) (5 loci).

Homoplasy occurring between divergent species is known to mask the true phylogeny and leads to inaccurate results. Inference of correct phylogenetic relationships using microsatellites is however possible at the intra-specific level (Schlotterer 2001), such as in this study among *M. balbisiana* genotypes and its sister species. Several observations indicated that this phylogeny produced accurate results among closely-related species. First, the closely-related species detected in chloroplast phylogeny such as *M. laterita*/*M. acuminata* subsp. *burmanica*; *M. basjoo*/*M. itinerans* and *M. manii*/*M. sanguinea*, remained closely-related in microsatellite phylogeny. Second, all *M. balbisiana* accession clustered in a single group. Third, previous studies on *Musa* genetic diversity found close genetic relationships between the four *M. balbisiana* accessions PKL, PKW, KTA and PBA, between the two accessions SIN and LVE (Carreel et al. 2002), and between the two accessions BUT and EKE (Ude et al. 2002a). These accessions remained closely-related in the microsatellite phylogeny. Fourthly, based on chloroplast and mitochondrial DNA RFLP, Carreel et al. in 2002 described five genetic groups using nine accessions of *M. balbisiana* species. Again, the same clustering of these accessions was observed with the microsatellite phylogeny. Finally, as expected between distant species, the basal branching order was poorly resolved and the nodes displayed low bootstrap values (< 50%), whereas the topology was well-resolved and more robust among *M. balbisiana* genotypes.

In our study, microsatellite phylogeny reflected the evolution of the whole nuclear genome. This was useful to construct phylogenetic relationships within *M. balbisiana* species. The first result concerned the analysis of the diversity and relatedness of *M. balbisiana*. Based on RFLP study, Carreel et al., 2002 (Carreel et al. 2002) found 5 genotypes of *M. balbisiana* among nine accessions tested (these accessions were included in the present study). The microsatellites analysis on a larger sampling of

M. balbisiana indicated that this species harbor a high genetic diversity. The microsatellite phylogeny confirmed the monophyly of *M. balbisiana*, and clarified the diversity and phylogenetic relationships between the 6 groups observed.

Additionally, we observed a correlation between the genetic differentiation of this species and the geographical pattern of the groups along a North-West to South-East transect of South Asia, where *M. balbisiana* originated. Group 3 contained accessions (SIN and LVE) collected in India, Group 4 contained accessions from Indonesia (PKL, PKW and PBA), Thailand (KTA) and Philippines (BUT), and Group 1 contained accessions from South-East Asia (HDN) and Papua-new Guinea (1016 and 626). The *M. balbisiana* species is supposed to originate from South China or India, which correspond to a major centre of diversification (Uma et al. 2005). It may then have reached eastern countries such as Thailand, Philippines, Northern Indonesia and Papua New Guinea via human settlements (Argent 1976; Cheesman 1947). Human populations used their fibers and consumed the unripe seeded-fruits and presumably introduced this species in eastern countries where it settled and became subsequently feral. Further population genetic studies coupled with data on migration events are required to investigate the biogeography of this species.

Both chloroplast and microsatellite phylogenies confirmed the paraphyly of *Rhodochlamys* section, also observed in AFLP and RFLP-based analysis (Nwakanma et al. 2003; Wong et al. 2002). Molecular phylogeny focused on *Eumusa* and *Rhodochlamys* should help to clarify the taxonomy.

Microsatellite phylogeny was also useful to assess the place of *M. boman*. Very few studies included *M. boman* and its phylogenetic position remained an enigma. First studies based on morphologic phylogenies placed *M. boman* close to *M. balbisiana* (Simmonds 1962) whereas others based on RFLP analysis placed this species in the section *Australimusa* (Gawel et al. 1992), a section well distinct from the *Eumusa/Rhodochlamys* clade. The microsatellite phylogeny finally strongly supported that *M. boman* belongs to the *Eumusa* section, and is a sister species of *M. balbisiana*.

For each integration studied, we designed several PCR markers specific for the internal rearrangements of EPRVs. As the complex structure of EPRV is assumed to arise at random via illegitimate recombination, we assumed that each structure is unique. Other PCR markers amplified the borders of EPRVs and enabled to detect homologous integration loci in the *Musa* genome. The same signature in different banana genomes would therefore reveal orthologous EPRV shared from a common ancestor.

The pattern of distribution of BSImV and BSGFV in *Musa* differed from each other. BSImV integration was restricted to the *M. balbisiana* genome whereas BSGFV is also integrated in the *M. boman* genome, as confirmed by southern blot hybridization. A first hypothesis would be that gene flow would have brought EPRV BSGFV from *M. balbisiana* into the *M. boman* genome. No data concerning *M. balbisiana* / *M. boman* hybridization exists so far, however interspecific hybrids between *M. balbisiana* and *M. acuminata* or *M. textilis* exist in natural populations (Simmonds 1962), as well as between *M. boman* and *M. lododensis* (*Australimusa* section) (Argent 1976). Furthermore, our chloroplast phylogeny might suggest the presence of gene flow between *M. boman* and species of section *Rhodochlamys*, also in line with this hypothesis. However, organelles and nuclear genomes do not necessarily share the same evolutionary history, and a more likely scenario can be proposed. This second hypothesis suggests that BSGFV integration occurred in the genome of the common ancestor of the two banana species, before the speciation *M. boman*/ *M. balbisiana*. BSImV EPRV would consequently be more recent than BSGFV EPRV. Since both EPRV are composed of repeated homologous regions which are good templates for intra-EPRV recombination, we expected a rapid evolution of the structure of EPRVs. This expectation was confirmed for EPRV BSImV only. EPRV BSGFV structure remained unaltered although it was older and had thus more time to accumulate mutations. These distinct evolutionary patterns between the two EPRVs might be explained by a distinct phenotypic consequence of their activation or by a distinct evolutionary dynamic of the genome at the integration site.

Concerning EPRV BSGFV, our results enabled to study the evolutionary dynamics of the infectious allele EPRV-7, and the non infectious allele EPRV-9. Early evolution of *M. balbisiana* suggests that the infectious allele appeared first, followed by evolution of allele EPRV-9. Since EPRV-9 differs from EPRV-7 from several substitutions and from a 2.3 kbp indel, multiple and independent evolution of a non infectious allele resembling to EPRV-9 from EPRV-7 is thus very unlikely. Additionally, the molecular detection of EPRV-9 was based on the presence of this particular indel, which implies that all EPRV-9 detected originate from a single event. EPRV-9 is nowadays present in all genetic groups (group Msat-1 to Msat-6) of *M. balbisiana*. We then expected that the non-infectious allele would have replaced the infectious one in wild *M. balbisiana* populations, but we observed the opposite. The infectious allele was kept and the non infectious allele was lost several times independently. To explain these results, we hypothesize that infectious EPRV might have played a beneficial role in the evolution of *M. balbisiana*, explaining why they were maintained in the host genome. *M. balbisiana* cv. PKW, and probably other genotypes of this species, evolved a resistance against EPRV activation and BSV infection after inoculation by the insect vector (mealybugs) (Iskra Caruana M.L. et al. 2003). Only the interspecific hybrids with *M. acuminata* are sensitive to EPRV activation. Infectious BSV EPRV would therefore be neutral for *M. balbisiana* plants, but be deleterious only in interspecific hybrids naturally occurring in sympatric (South-East Asia) (Simmonds 1962). Synthetic hybrids from contemporaneous *M. balbisiana* x *M. acuminata* are polyploid and poorly fertile and inbreeding depression would have prevented large introgression of nuclear DNA between these two species in natural populations (Simmonds 1962). Infectious EPRVs might have also contributed to reinforce the speciation of *M. balbisiana* and *M. acuminata* species by producing less fit BSV-infected hybrids.

Large scale phylogenetic analysis conducted on different episomal and integrated BSV species confirmed that the majority of integrations in the *Musa* genomes occurred after the speciation between *M. acuminata* and *M. balbisiana* (Geering et al. 2005; Gayral al, unpublished data). The recent origin of EPRV integrations was confirmed in this study by the analysis of substitution rate of the BSGFV EPRV locus.

Our results suggest this integration could have occurred approximately 0.63 MYA, i.e. recently and long after the speciation event estimated at 4.5 MYA (Lescot et al. 2008). Allelic divergence between the infectious and the non-infectious BSGFV allele is much more recent since it may have arisen 0.23 MYA. Our results suggest that BSV EPRVs are recent, still displaying functional coding and regulatory viral sequences found in infectious EPRVs. Figure 6 summarizes the evolutionary history events of BSV EPRVs and their host *Musa*. We must emphasize that all of these divergence dates are approximate, based on rough estimates of minimum divergence times. However, this approach has been successfully used for transposable elements in an other plant (SanMiguel et al. 1998) and tends to underestimate the divergence time, since divergence could be diminished by gene conversion occurring between homologous sequences.

Our study finally enabled to retrace the phylogenetic relationships between integrated sequences and episomal viruses collected on domesticated bananas carrying the B genome. We found little variations between BSImV sequences, either from integrated or episomal origin. This homogeneity is presumably the result of insufficient sampling. Additional sequence data are therefore needed to produce a more representative sampling of the genetic diversity of episomal BSImV species. Alternatively, this situation could indicate that episomal BSImV sampled in Colombian and Ugandan banana crops both derived from recent activation of the same EPRV. Infectious EPRV would therefore increase the inclusive fitness of closely related viruses, and would prevent viral populations from local extinctions. Concerning BSGFV species, we confirmed the presence of three distinct genetic groups as previously described (Harper et al. 2005), and showed that EPRV derived from Group 1 viruses. Interestingly, episomal 'BSGFV' sequence - corresponding to the complete BSGFV genomic sequence - clustered near EPRV clade, and was primarily isolated from interspecific hybrid containing B genome. These results suggest that this episomal virus could have been restituted from infectious EPRVs. Furthermore, we expected that this episomal virus accumulated more mutations than EPRV sequences because BSV genome evolved quicker than those of *Musa* (Gayral al,

unpublished data). This trend was confirmed by observing longer branch for episomal sequences than in endogenous sequences.

This study showed that endogenous BSGFV and BSImV sequences were useful to retrace the recent history of infectious EPRVs. They were however too recent and therefore not enough polymorphic to be used as phylogenetic markers of *M. balbisiana*. Further investigation on other BSV EPRV that integrated before the speciation between *M. acuminata* and *M. balbisiana* (Gayral al, unpublished data), estimated to have occurred 4.5 My ago (Lescot et al. 2008), would be needed for their evaluation as good phylogenetic markers of *Musa* species.

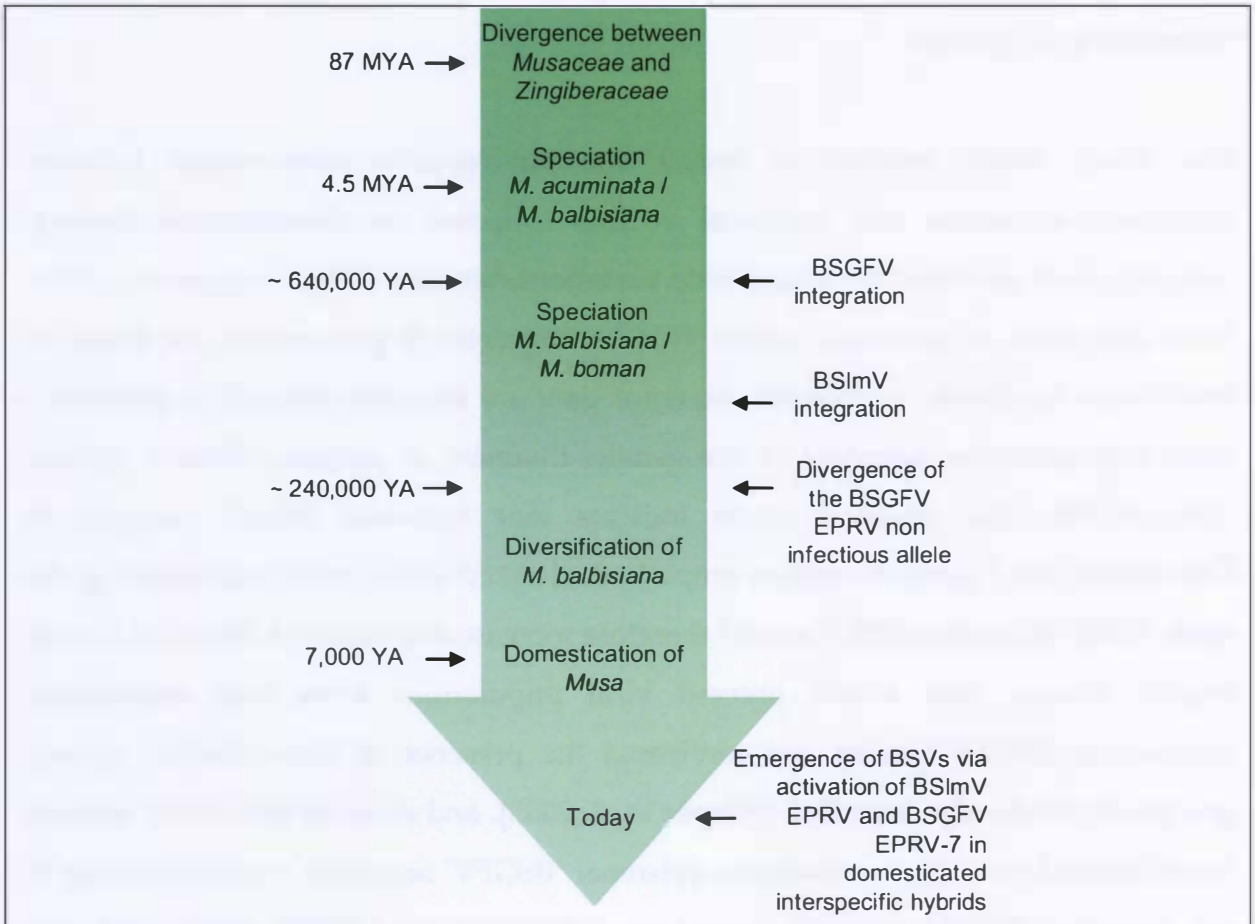

Figure 6: Dates and events associating the evolutionary history of BSV EPRVs and their host. Schematic diagram illustrating key events of the evolutionary history of BSImV and BSGFV described in this study in association with estimated dates during the evolution of the *Musa* genus. Estimations of the age of the most recent common ancestor for the *Musacaceae* and *Zingiberaceae* are from Bremer (2000); Kress et al. (2001); Sanderson et al. (2004), and between *M. acuminata* / *M. balbisiana* from Lescot et al. (2008). Date of emergence of agriculture and domestication of banana are from Denham et al. (2003).

Acknowledgements

We are very grateful to Serge Galzi and Nathalie Laboureau for technical assistance, to Liying Zhang and Benham E. L. Lockhart for providing the two BSGFV clones and Andrew Geering for performing the sequencing of BSImV genome. We thank the curators of *Musa* collections for providing vitroplants and leaves samples: Christophe Jenny (Station de Recherches Fruitières de Neufchâteau, CIRAD), Ines Van Den Houwe and A. M. Ayodele (INIBAP Transit Center (ITC), Katholieke Universiteit Leuven), Perpetua Udu and A. Tenkouano (International Institute of Tropical Agriculture (IITA ONNE), Onne Station). We thank Elisabeth Fournier and Didier Tharreau for helpful comments. Materials and methods fulfilled the requirements of quality management system according to the guidelines of ISO 9001:2000 standard. P. G. was supported by a PhD grant 'CIRAD-Région Languedoc Roussillon'.

Literature cited

Argent GCG. 1976. The wild bananas of Papua New Guinea. Notes Roy Bot Gard Edinburgh 35(1):77 - 114.

Bennetzen JL. 2000. Transposable element contributions to plant gene and genome evolution. Plant Mol Biol 42(1):251-269.

Bousalem M, Douzery EJP, Seal SE. 2008. Taxonomy, molecular phylogeny and evolution of plant reverse transcribing viruses (family Caulimoviridae) inferred from full-length genome and reverse transcriptase sequences. Arch Virol 153(6):1085-1102.

Bremer K. 2000. Early Cretaceous lineages of monocot flowering plants. Proceedings of the National Academy of Sciences of the United States of America 97(9):4707-4711.

Carreel F, de Leon DG, Lagoda P, Lanaud C, Jenny C, Horry JP, du Montcel HT. 2002. Ascertaining maternal and paternal lineage within *Musa* by chloroplast and mitochondrial DNA RFLP analyses. Genome 45(4):679-692.

Cheesman EE. 1947. Classification of the banana. Kew bulletin 2:97-117.

Chirhart SE, Honeycutt RL, Greenbaum IF. 2005. Microsatellite variation and evolution in the *Peromyscus maniculatus* species group. Mol Phyl Evol 34(2):408-415.

Cuenoud P, Savolainen V, Chatrou LW, Powell M, Grayer RJ, Chase MW. 2002. Molecular phylogenetics of Caryophyllales based on nuclear 18S rDNA and plastid rbcL, atpB, and matK DNA sequences. Am J Bot 89(1):132-144.

Dallot S, Acuna P, Rivera C, Ramirez P, Cote F, Lockhart BEL, Caruana ML. 2001. Evidence that the proliferation stage of micropropagation procedure is determinant in the expression of *Banana streak virus* integrated into the genome of the FHIA 21 hybrid (*Musa* AAAB). Arch Virol 146(11):2179-2190.

Denham TP, Haberle SG, Lentfer C, Fullagar R, Field J, Therin M, Porch N, Winsborough B. 2003. Origins of agriculture at Kuk Swamp in the highlands of New Guinea. Ann Bot (Lond) 301(5630):189-193.

Duffy S, Shackelton LA, Holmes EC. 2008. Rates of evolutionary change in viruses: patterns and determinants. Nat Rev Genet 9(4):267-276.

Fargette D, Konate G, Fauquet C, Muller E, Peterschmitt M, Thresh JM. 2006. Molecular ecology and emergence of tropical plant viruses. Annu Rev Phytopathol 44:235-260.

Faure S, Noyer JL, Carreel F, Horry JP, Bakry F, Lanaud C. 1994. Maternal inheritance of chloroplast genome and paternal inheritance of mitochondrial genome in bananas (*Musa acuminata*). Current genetics 25(3):265-269.

Galtier N, Gouy M, Gautier C. 1996. SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. Comput Appl Biosci 12(6):543-548.

Gawel NJ, Jarret RL, Whittemore AP. 1992. Restriction fragment length polymorphism (RFLP)-based phylogenetic analysis of *Musa*. Theor Appl Genet 84(3):286-290.

Gawel NJ, Jarret, R.L. 1991. A modified CTAB DNA extraction procedure for Musa and Ipomoea. Plant Mol Biol Rep 9:262-266.

Gayral P, Noa-Carrazana J-C, Lescot M, Lheureux F, Lockhart BEL, Matsumoto T, Piffanelli P, Iskra-Caruana M-L. 2008. A single *Banana streak virus* integration event in the banana genome as the origin of infectious endogenous pararetrovirus. J Virol 82(13):6697-6710.

Ge XJ, Liu MH, Wang WK, Schaal BA, Chiang TY. 2005. Population structure of wild bananas, *Musa balbisiana*, in China determined by SSR fingerprinting and cpDNA PCR-RFLP. Mol Ecol 14(4):933-944.

Geering ADW, Olszewski NE, Harper G, Lockhart BEL, Hull R, Thomas JE. 2005. Banana contains a diverse array of endogenous badnaviruses. J Gen Virol 86:511-520.

Goudet J. 1995. FSTAT (Version 1.2): a computer program to calculate F-statistics. J Hered 86(6):485-486.

Gregor W, Mette MF, Staginnus C, Matzke MA, Matzke AJM. 2004. A distinct endogenous pararetrovirus family in *Nicotiana tomentosiformis*, a diploid progenitor of polyploid tobacco. Plant Physiol 134(3):1191-1199.

Guindon S, Lethiec F, Duroux P, Gascuel O. 2005. PHYML Online - a web server for fast maximum likelihood-based phylogenetic inference. Nucleic acids research 33(Web Server issue):W557-559.

Hall TA. 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucl Acids Symp Ser 41:95-98.

Harper G, Hart D, Moult S, Hull R, Geering A, Thomas J. 2005. The diversity of *Banana streak virus* isolates in Uganda. Arch Virol:-46.

Harr B, Weiss S, David JR, Brem G, Schlötterer C. 1998. A microsatellite-based multilocus phylogeny of the *Drosophila melanogaster* species complex. Curr Biol 8(21):1183-1187.

Heslop-Harrison JS, Schwarzacher T. 2007. Domestication, genomics and the future for banana. Ann Bot (Lond).

Hohn T, Richert-Poggeler KR, Harper G, Schwarzacher T, Teo CH, Techeney PY, Iskra-Caruana ML, Hull R. 2008. Evolution of integrated plant viruses. In Virus Evolution. Springer, Heidelberg, M. Roosinck ed. In Press.

Huelsenbeck JP, Ronquist F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17(8):754-755.

Hull R. 1999. Classification of reverse transcribing elements: a discussion document. Arch Virol 144(1):209-214.

Hull R, Covey SN. 1995. Retroelements: propagation and adaptation. Virus Genes 11(2-3):105-118.

Iskra Caruana M.L., Lheureux F., Noa-Carrazana J.C., Piffanelli P., Carreel F., Jenny C., Laboureau N., Lockhart B.E.L. 2003. Unstable balance of relation between pararetrovirus and its host plant: the BSV-EPRV banana pathosystem, p. 8. EMBO Workshop: Genomic Approaches in Plant Virology, Keszthely, Hungary.

Jakowitsch J, Mette MF, van der Winden J, Matzke MA, Matzke AJM. 1999. Integrated pararetroviral sequences define a unique class of dispersed repetitive DNA in plants. Proc Natl Acad Sci USA 96(23):13241-13246.

Kaemmer D, Fischer D, Jarret RL, Baurens FC, Grapin A, Dambier D, Noyer JL, Lanaud C, Kahl G, Lagoda PJL. 1997. Molecular breeding in the genus *Musa*: a strong case for STMS marker technology. Euphytica 96(1):49-63.

Kidwell MG, Lisch DR. 2000. Transposable elements and host genome evolution. Trends Ecol Evol 15(3):95-99.

Kress WJ, Prince LM, Hahn WJ, Zimmer EA. 2001. Unraveling the evolutionary radiation of the families of the Zingiberales using morphological and molecular evidence. Syst Biol 50(6):926-944.

Lagoda PJ, Noyer JL, Dambier D, Baurens FC, Grapin A, Lanaud C. 1998. Sequence tagged microsatellite site (STMS) markers in the *Musaceae*. Mol Ecol 7(5):659-663.

Lescot M, Piffanelli P, Ciampi AY, Ruiz M, Blanc G, Leebens-Mack J, da Silva FR, Santos CM, D'Hont A, Garsmeur O, Vilarinhos AD, Kanamori H, Matsumoto T, Ronning CM, Cheung F, Haas BJ, Althoff R, Arbogast T, Hine E, Pappas GJ, Sasaki T, Souza MT, Miller RN, Glaszmann JC, Town CD. 2008. Insights into the *Musa* genome: syntenic relationships to rice and between *Musa* species. BMC Genomics 9(1):58.

Lheureux F, Carreel F, Jenny C, Lockhart B, Iskra-Caruana M. 2003. Identification of genetic markers linked to banana streak disease expression in inter-specific *Musa* hybrids. Theor Appl Genet 106(4):594-598.

Lheureux F, Laboureau N, Muller E, Lockhart BE, Iskra-Caruana ML. 2007. Molecular characterization of *Banana streak acuminata Vietnam virus* isolated from *Musa acuminata siamea* (banana cultivar). Arch Virol 152(7):1409-1416.

Lockhart B, Jones D. 2000. Banana streak. In diseases of banana, abaca and enset, ed. DR Jones, pp. 263–74. Wallingford, UK CAB Int.

Lockhart BE, Menke J, Dahal G, Olszewski NE. 2000. Characterization and genomic analysis of *Tobacco vein clearing virus*, a plant pararetrovirus that is transmitted vertically and related to sequences integrated in the host genome. J Gen Virol 81:1579-1585.

Malik HS, Eickbush TH. 2001. Phylogenetic analysis of ribonuclease H domains suggests a late, chimeric origin of LTR retrotransposable elements and retroviruses. Genome Res 11(7):1187-1197.

Matzke M, Gregor W, Mette MF, Aufstatz W, Kanno T, Jakowitsch J, Matzke AJM. 2004. Endogenous pararetroviruses of allotetraploid *Nicotiana tabacum* and its diploid progenitors, *N. sylvestris* and *N. tomentosiformis*. Biol J Linn Soc 82(4):627-638.

Muir G, Fleming CC, Schlotterer C. 2000. Species status of hybridizing oaks. Nature 405(6790):1016.

Ndowora T, Dahal G, LaFleur D, Harper G, Hull R, Olszewski NE, Lockhart B. 1999. Evidence that *Badnavirus* infection in *Musa* can originate from integrated pararetroviral sequences. Virology 255(2):214-220.

Noreen F, Akbergenov R, Hohn T, Richert-Poggeler KR. 2007. Distinct expression of endogenous *Petunia vein clearing virus* and the DNA transposon dTph1 in two *Petunia hybrida* lines is correlated with differences in histone modification and siRNA production. Plant J 50(2):219-229.

Nwakanma DC, Pillay M, Okoli BE, Tenkouano A. 2003. Sectional relationships in the genus *Musa* L. inferred from the PCR-RFLP of organelle DNA sequences. Theor Appl Genet 107(5):850-856.

Pahalawatta V, Druffel K, Pappu H. 2008. A new and distinct species in the genus *Caulimovirus* exists as an endogenous plant pararetroviral sequence in its host, *Dahlia variabilis*. Virology 376(2):253-257.

Richard M, Thorpe RS. 2001. Can microsatellites be used to infer phylogenies? Evidence from population affinities of the Western Canary Island lizard (*Gallotia galloti*). Mol Phyl Evol 20(3):351-360.

Richert-Poggeler KR, Noreen F, Schwarzacher T, Harper G, Hohn T. 2003. Induction of infectious petunia vein clearing (pararetro) virus from endogenous provirus in petunia. Embo J 22(18):4836-4845.

Ritz LR, Glowatzki-Mullis ML, MacHugh DE, Gaillard C. 2000. Phylogenetic analysis of the tribe *Bovini* using microsatellites. Animal genetics 31(3):178-185.

Sanderson MJ, Thorne JL, Wikstrom N, Bremer K. 2004. Molecular evidence on plant divergence times. Am J Bot 91(10):1656-1665.

SanMiguel P, Gaut BS, Tikhonov A, Nakajima Y, Bennetzen JL. 1998. The paleontology of intergene retrotransposons of maize. Nature genetics 20(1):43-45.

Schlotterer C. 2001. Genealogical inference of closely related species based on microsatellites. Genet Res 78(3):209-212.

Simmonds NW, editor. 1962. The evolution of the bananas, Longmans Green (Ed). London. p 170.

Simmonds NW, Weatherup STC. 1990. Numerical taxonomy of the wild bananas (*Musa*). pp 567-571.

Staginnus C, Gregor W, Mette MF, Teo CH, Borroto-Fernandez EG, Machado ML, Matzke M, Schwarzacher T. 2007. Endogenous pararetroviral sequences in tomato (*Solanum lycopersicum*) and related species. BMC Plant Biol 7:24.

Staginnus C, Richert-Poggeler KR. 2006. Endogenous pararetroviruses: two-faced travelers in the plant genome. Trends Plant Sci 11(10):485-491.

Su L, Gao S, Huang Y, Ji C, Wang D, Ma Y, Fang R, Chen X. 2007. Complete genomic sequence of *Dracaena mottle virus*, a distinct *Badnavirus*. Virus Genes 35(2):423-429.

Taberlet P, Gielly L, Pautou G, Bouvet J. 1991. Universal primers for amplification of three non-coding regions of chloroplast DNA. Plant Mol Biol 17(5):1105-1109.

Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucl Acids Res 22(22):4673-4680.

Ude G, Pillay M, Nwakanma D, Tenkouano A. 2002a. Analysis of genetic diversity and sectional relationships in *Musa* using AFLP markers. Theor Appl Genet 104(8):1239-1245.

Ude G, Pillay M, Nwakanma D, Tenkouano A. 2002b. Genetic Diversity in *Musa acuminata* Colla and *Musa balbisiana* Colla and some of their natural hybrids using AFLP Markers. Theor Appl Genet 104(8):1246-1252.

Uma S, Siva SA, Saraswathi MS, Durai P, Sharma T, Singh DB, Selvarajan R, Sathiamoorthy S. 2005. Studies on the origin and diversification of Indian wild banana (*Musa balbisiana*) using arbitrarily amplified DNA markers. J Hortic Sci Biotech 80(5):575-580.

Wong C, Kiew R, Argent G, Set O, Lee SK, Gan YY. 2002. Assessment of the validity of the sections in *Musa* (Musaceae) using AFLP. Ann Bot-London 90(2):231-238.

Xia X. 1999. DAMBE (Software Package for Data Analysis in Molecular Biology and Evolution). User Manual. Hong Kong: Department of Ecology and Biodiversity, University of Hong Kong.

Xia X, Xie Z. 2001. DAMBE: software package for data analysis in molecular biology and evolution. J Hered 92(4):371-373.

Xia X, Xie Z, Salemi M, Chen L, Wang Y. 2003. An index of substitution saturation and its application. Molecular Phylogenetics and Evolution 26(1):1-7.

CONCLUSION GENERALE

L'ère de la génomique a permis de révéler ces dernières années l'importance, du point de vue quantitatif, des intégrations de *Caulimoviridae* dans le génome des plantes, en particulier chez les solanacées et les musacées. Ces découvertes nous ont éclairés de manière globale sur les genres viraux intégrés, sur la quantité d'EPRV par génome, ou sur le fait que les EPRV de certaines plantes sont régulés par des mécanismes épigénétiques.

Cependant chaque modèle d'EPRV apparaît complexe et multiple. Les EPRV de plusieurs espèces virales peuvent en effet coexister au sein d'un même génome, et des EPRV *a priori* similaires peuvent avoir subi des histoires évolutives différentes. De plus, les EPRV infectieux ont des effets similaires à ceux induits par les virus dont ils proviennent. Des études fines et spécifiques sont donc nécessaires pour définir leurs particularités biologiques et préciser leur histoire évolutive.

Ce travail de thèse s'inscrit clairement dans cette démarche. Il vise à mieux comprendre le phénomène d'intégration des *Caulimoviridae* chez les plantes, en utilisant les EPRV du BSV comme modèle biologique. Pour y parvenir, nous avons d'abord tenté de comprendre le fonctionnement d'une intégration infectieuse particulière : les EPRV du *Banana streak GF virus* intégrés chez le bananier *Musa balbisiana* cv. PKW. Nous avons ensuite étudié l'évolution des EPRV à deux niveaux d'échelle. Une première étude à large échelle nous a tout d'abord permis d'observer les tendances globales de l'évolution moléculaire des EPRV et des intégrations des différentes espèces de BSV chez plusieurs espèces de bananiers. Nous avons ensuite conduit une étude plus fine pour retracer l'histoire évolutive de deux EPRV BSV pathogènes des espèces BSGFV et BSImV au sein du genre *Musa*.

Conclusion

Phénomène d'intégration des EPRV

Conduite préalablement au Cirad, l'analyse par Fingerprint des clones BAC du génome diploide *M. balbisiana* contenant des EPRV BSGFV a révélé la présence de

deux types d'intégrants dans le cv. PKW. De manière similaire à ce qui a déjà été observé pour l'EPRV BSOLV étudié en 1999 par Ndowora et al., les EPRV BSGFV ont une structure fortement remaniée. Le locus d'intégration est composé de plusieurs fragments tronqués du génome viral, orientés dans les deux sens. Ces remaniements ont conduit, chez le cv. PKW, à la formation de deux types d'EPRV, EPRV-7 et EPRV-9. L'analyse des gènes *Musa* présents de part et d'autre de ces EPRV suggèrent que les intégrants étaient des formes alléliques (travaux préliminaires de M. Lescot à l'UMR DAP). Le génotypage de chaque EPRV a permis l'étude de leur ségrégation dans la descendance du croisement interspécifique utilisant le cv. PKW comme parent *M. balbisiana*. Nous avons montré que ces deux EPRV sont deux allèles du même locus et résultent d'un unique évènement d'intégration dans le génome du bananier étudié.

L'analyse plus précise de l'environnement génétique de l'EPRV BSGFV a montré que l'intégration est associée à un rétrotransposon à LTR de type Ty3/Gypsy. Une telle association, EPRV-rétroélément, est fréquemment observée dans de nombreux modèles d'EPRV des solanacées (Gregor et al., 2004; Jakowitsch et al., 1999; Richert-Pöggeler et al., 2003; Staginnus et al., 2007), et également rapportée pour le bananier Obino l'Ewai (Ndowora et al., 1999). La particularité de l'EPRV BSGFV décrite dans notre étude et qui n'avait jamais été observée, est une intégration du BSV à l'intérieur d'un rétrotransposon complet. Cette intégration peut résulter d'une recombinaison non homologue au niveau de microhomologies entre les deux séquences, ou de la rétrotransposition du rétroélément après recombinaison entre les ARN prégénomiques viraux et les ARN du rétroélément en activité. Cette dernière hypothèse est celle retenue pour expliquer les intégrations des *Potyviridae* dans la vigne et des *Discistroviridae* dans le génome des abeilles (Maori *et al.*, 2007; Tanne & Sela, 2005).

Dans le cas où ce rétroélement est encore actif, il n'est pas exclu qu'à plus long terme évolutif, des stress favorisant la réactivation des rétroéléments puissent amplifier l'EPRV BSGFV au sein du génome B, comme cela a été observé pour les EPRV des Solanacées (Matzke et al., 2004; Richert-Pöggeler et al., 2003).

Deuxièmement, cette chimère EPRV-rétrotransposon est située à l'intérieur d'un intron du gène *mom*, pseudogène à l'heure actuelle. Aucun autre cas d'intégration de *Caulimoviridae* dans un gène de l'hôte n'a été rapporté dans la littérature. Les EPRV des Solanacées (genres *Petunia*, *Nicotiana* et *Solanum*), sont préférentiellement liés à l'hétérochromatine (Hansen et al., 2005; Richert-Pöggeler et al., 2003; Staginnus et al., 2007), qui correspond à des zones chromosomiques plutôt pauvres en gènes et riches en rétroéléments. Les conséquences de la présence de l'EPRV BSGFV dans le gène *mom* de *M. balbisiana* sont à l'heure actuelle inconnues, mais il est troublant de noter que l'homologue de ce gène est impliqué dans le contrôle épigénétique des éléments répétés chez *A. thaliana* (Amedeo et al., 2000; Vaillant et al., 2006). Nous pouvons supposer que ce gène ayant perdu sa fonction suite à l'intégration du BSGFV, ne régulerait plus les rétroéléments et les EPRV infectieux des autres espèces virales de BSV, qui seraient alors transcrits et activés.

Structure des EPRV et mécanismes d'activation.

Depuis les travaux de Ndowora et co-auteurs en 1999 sur l'EPRV BSOLV chez le cv. Obino l'Ewai, aucun autre exemple d'EPRV infectieux du BSV n'a été reporté dans la littérature. Nous avons démontré qu'un deuxième EPRV du BSV appartenant à l'espèce BSGFV est présent dans le génome des *M. balbisiana*, et qu'il est infectieux.

Le cv. PKW est porteur sain et toujours indemne de virions, et l'activation des EPRV s'observe seulement dans la descendance hybride interspécifique triploïde (AAB) impliquant le cv. PKW comme parent *M. balbisiana*. L'analyse de l'activation du BSV dans la descendance a montré qu'un seul des deux allèles (EPRV-7) est infectieux puisque seuls les hybrides porteurs de cet allèle sont capables de restituer un virus BSGFV fonctionnel.

Les régions codantes de l'allèle non-infectieux (EPRV-9) contiennent des mutations délétères pouvant expliquer le fait qu'aucun virus fonctionnel ne puisse être restitué à partir de cet allèle. L'allèle infectieux, quant à lui, possède des ORF indemnes de mutations fortement délétères. De manière troublante, la formation d'un génome

viral à partir de l'EPRV ne peut se faire directement. Ce n'est qu'après une série de modifications de la structure physique de l'EPRV que le génome viral fonctionnel est restitué. Nous avons reconstruit un modèle théorique d'activation basé sur des recombinaisons homologues. Selon ce modèle, deux étapes de recombinaison au minimum sont requises pour produire un génome viral fonctionnel, c'est-à-dire complet et circulaire, à partir de l'allèle infectieux du BSGFV. La validation de ce modèle s'est faite en deux temps. D'une part, nous avons vérifié que les zones de l'EPRV impliquées dans la restitution du génome viral ne contenaient pas de mutations fortement délétères. D'autre part, nous avons vérifié que les molécules recombinées correspondant aux étapes intermédiaires ainsi qu'aux produits finaux étaient détectées dans le modèle bactérien utilisé, ainsi que chez le cv. PKW. Ces molécules sont probablement en très faible quantité, à la limite de la détection par PCR, ce qui expliquerait qu'une des quatre régions recombinées n'ait pas été observée chez le cv. PKW. Enfin, il est important de noter que la recombinaison homologue, bien que rare chez les plantes, n'est pas le facteur limitant de l'activation des EPRV, puisqu'elle s'observe sans induction particulière *in planta*, et qu'elle semble également concerner l'allèle non infectieux. A l'opposé, la présence d'un génome intégré complet possédant des ORF non altérées semble être un facteur nécessaire à la multiplication virale.

La présence de l'allèle infectieux EPRV-7 est une condition nécessaire à l'activation, mais elle n'est pas suffisante. En effet, la majorité (75%) des hybrides porteurs de l'allèle infectieux ne sont pas infectés par le virus BSGFV. Une explication serait que des facteurs génétiques ou épigénétiques, combinés à des facteurs environnementaux jouent également un rôle dans l'issue des infections initiées lors de l'activation d'un EPRV infectieux.

L'activation apparaît être un phénomène qui comporte plusieurs étapes. La première est la restitution d'un génome viral infectieux, par recombinaison homologue dans le cas de l'allèle EPRV-7 du BSGFV. Cependant, des génomes non infectieux et incapables de se répliquer peuvent également être générés lors de cette première étape, à partir de l'allèle EPRV-9 du BSGFV intégré par exemple. La deuxième étape concerne le début l'infection, lors de l'initiation du cycle viral à partir des quelques

génomes infectieux présents dans les tissus de la plante. Une fois qu'un génome viral infectieux a été produit, les plantes peuvent se défendre contre l'infection débutante par l'expression de gènes de défenses, ainsi que par des mécanismes de RNAi activés par la transcription des EPRV et/ou par les ARN viraux.

Des intégrations récentes et simultanées du BSV

Nous avons souhaité dans un premier temps apporter des clarifications sur les relations phylogénétiques des BSV, au travers d'une analyse phylogénétique conduite sur les séquences virales intégrées et libres. Ainsi, cette étude a pour la première fois pris en compte les informations apportées par les EPRV pour les intégrer à la phylogénie des séquences des virus libres. La prise en compte de toutes les séquences proches du BSV alors disponibles dans les banques de données nous a permis d'étudier à grande échelle le phénomène d'intégration.

Tout d'abord, un nombre important d'intégrations virales indépendantes se sont produites dans les trois espèces de bananiers étudiées, faisant de l'intégration un phénomène fréquent. Ensuite, la majorité de ces intégrations se sont produites après la spéciation *M. acuminata* et *M. balbisiana*, estimée à 4,5 Ma, c'est-à-dire récemment d'un point de vue évolutif. L'âge exact des intégrations reste toutefois inconnu. Il est cependant possible d'estimer celui de l'EPRV BSGFV étudié dans cette thèse et de l'utiliser comme ordre de grandeur pour estimer celui des autres EPRV. Ainsi, si l'on considère que l'EPRV est arrivé avec le rétroélément (hypothèse I présentée dans l'article 1), nous pouvons alors utiliser la date approximative de 0,63 Ma calculée à partir de la divergence des LTR et du taux de substitution des *Musaceae* (Lescot et al., 2008), comme date d'intégration de l'EPRV BSGFV. Bien que ce calcul soit très approximatif, il semble confirmer que les intégrations sont récentes par rapport à l'émergence de l'espèce *M. balbisiana*. Ces intégrations seraient donc relativement jeunes d'un point de vue évolutif, ce qui pourrait expliquer la présence de séquences virales toujours fonctionnelles que l'on trouve dans les EPRV infectieux.

Evolution des EPRV

L'estimation des paramètres évolutifs des séquences du BSV a mis en évidence que les différences observées d'évolution moléculaire entre les virus épisomaux et les séquences intégrées, étaient bien moindres que celles attendues théoriquement. Ce résultat suggèrerait que les EPRV participent activement aux infections et modifient significativement les paramètres évolutifs des BSV, et/ou que les intégrations des BSV sont des phénomènes fréquents et réguliers, puisqu'ils s'observent également sur les branches terminales des arbres phylogénétiques.

La construction d'une phylogénie des bananiers *M. balbisiana* nous a permis d'interpréter la distribution des EPRV infectieux de deux espèces de BSV : BSGFV et BSImV. L'intégration unique de BSGFV s'est probablement produite avant la spéciation *M. balbisiana* et *M. boman*, et serait antérieure à celle de BSImV qui n'est observée que chez les *M. balbisiana*. L'EPRV BSGFV s'est intégré sous la forme de l'allèle infectieux, l'allèle non infectieux ayant divergé ensuite. La date de cette divergence peut être estimée (sachant le taux de substitution entre allèles et la vitesse d'évolution du génome *Musa*) à approximativement 0.23 Ma, soit 0.4 Ma après l'intégration. L'allèle non infectieux n'a par la suite pas été fixé chez les *M. balbisiana*, probablement à cause de la dérive génétique, ou du fait qu'il ne représente pas réellement un désavantage sélectif pour les bananiers *M. balbisiana* (voir plus bas).

Conséquences évolutives des EPRV pour les bananiers

Il n'est pas rare qu'un phénomène accidentel puisse avoir des conséquences majeures pour l'évolution des organismes. C'est le cas des EPRV chez les plantes, et chez les bananiers en particulier en provoquant l'apparition du BSV à partir du génome d'une plante saine.

Deux conséquences majeures liées à la présence des EPRV infectieux peuvent être observées. La première est associée à la résistance de *M. balbisiana* envers l'activation des EPRV. Nous avons montré que dans le cas de l'EPRV BSGFV du moins, cette résistance semble empêcher la multiplication des BSV dans le cv. PKW, plutôt que cibler les étapes précoces de l'activation comme par exemple une diminution de la

recombinaison homologue. Nous pouvons penser que le maintien d'un système de défense contre le 'danger permanent' que sont les EPRV infectieux du BSV, entraîne vraisemblablement un coût pour les bananiers *M. balbisiana.*

La deuxième conséquence de la présence d'EPRV infectieux est le fait que certains génotypes de bananiers sont sensibles à l'infection du BSV provoquée par l'activation des EPRV. Il est important de noter qu'à ce jour, les seuls bananiers concernés par le réveil des EPRV infectieux sont les hybrides interspécifiques entre les espèces *M. balbisiana* et *M. acuminata* (génotypes AAB ou AAAB). C'est ce que nous avons démontré lors de l'étude des EPRV BSGFV chez les hybrides interspécifiques AAB. D'un point de vue évolutif, les EPRV infectieux n'ont donc pas de conséquences évolutives néfastes pour les bananiers *M. balbisiana* (si ce n'est un hypothétique coût de la résistance), puisque la réduction de fitness associée aux EPRV concerne les hybrides interspécifiques seulement, et non les individus *M. balbisiana.* Par conséquent, les EPRV infectieux ont peut-être pu constituer une force évolutive qui aurait participé au renforcement de l'isolement reproductif des espèces *M. balbisiana* et *M. acuminata* que l'on observe aujourd'hui, puisque ces hybrides sont en général stériles.

EPRV : un nouveau type de parasites ?

Les EPRV ont souvent été qualifiés de parasites, puisqu'ils peuvent être infectieux pour les plantes et qu'ils participent dans une mesure importante à l'émergence du virus dans les cultures. Cependant, les intégrations sont accidentelles, et la sélection naturelle ne joue pas en faveur du maintien des capacités codantes ou infectieuses des EPRV. En d'autres termes, les EPRV à eux seuls ne peuvent pas maintenir leur intégrité physique dans le temps. De ce point de vue strict, les EPRV infectieux ne peuvent pas être considérés comme des organismes. Par contre, si l'on considère l'information génétique comme l'entité biologique transmise, alors les EPRV sont des parasites à part entière, puisqu'ils sont capables pendant un certain temps, de reproduire les gènes du virus. Bien que les intégrations ne semblent pas être issus

d'une stratégie virale, les EPRV infectieux peuvent donc être vus comme un 'génotype étendu' des virus épisomaux correspondants.

Perspectives :

EPRV et mutations d'insertion ?

A long terme :

L'analyse de l'environnement génomique de l'EPRV BSGFV a révélé que l'intégration s'est produite dans un intron de pseudogène homologue au gène *mom*. Nous proposons de tester si nous avons à faire à un cas de mutation d'insertion causée par l'EPRV. Il est en effet probable que l'intégration virale soit responsable de la perte de fonction de ce gène. De manière surprenante, le gène mom d'*Arabidopsis thaliana* participe au contrôle épigénétique des éléments répétés du génome. Il serait alors intéressant d'étudier la fonction de ce gène chez le bananier, afin de vérifier si sa pseudogénisation est associée à des changements de nature épigénétique, qui pourraient à leur tour participer à l'activation des autres EPRV du génome. Il pourrait être envisageable d'étudier ce gène chez des bananiers possédant le gène *mom* fonctionnel, peut-être chez *M. acuminata*. Etant donné que la transformation stable des bananiers est une entreprise périlleuse car l'obtention de bananiers adultes est longue, l'expression du gène *mom* pourrait être réduite via des constructions RNAi transformées soit par agroinfiltration (Kaku et al., 2006), soit par biolistique (Douchkov et al., 2005) sur des cultures cellulaires.

Mécanismes d'activation des EPRV BSV

A court terme :

Le modèle d'activation des EPRV basé sur des étapes de recombinaison homologue fut proposé initialement en 1999 par Ndowora et co-auteurs, puis relayé dans la littérature. Dans notre étude, la structure des EPRV BSGFV ressemble à celle du BSOLV de Ndowora et al. (1999), et un modèle de recombinaison homologue semble être le plus adapté aux données. Cependant, il n'est pas exclu que pour d'autres EPRV infectieux du génome B, d'autres mécanismes soient responsables de

l'activation. Lorsque la structure des EPRV est un génome entier présentant des répétitions en tandem de génomes viraux, la transcription directe apparait alors comme un mécanisme plus probable pour produire des ARN prégénomiques (cas du PVCV du pétunia). Il faudra donc connaitre les séquences des autres EPRV infectieux du génome du cv. PKW (projet de séquençage et annotation des clones BAC par le Génoscope en phase finale), avant de proposer un mécanisme d'activation par transcription plutôt que par recombinaison homologue.

Concernant les EPRV BSGFV, un modèle mixte combinant recombinaison puis transcription peut également être envisagé. Il sera également nécessaire de tester expérimentalement ce modèle mixte, en recherchant des transcrits chez le cv. PKW qui correspondent à la molécule recombinée après une première étape de recombinaison homologue.

A court terme :

Un projet de bombardement d'EPRV clonés a été initié lors du stage de master 2 d'Olivier Guidolin. Un des objectifs de ce projet était d'apporter une démonstration formelle de la nature infectieuse des EPRV, en introduisant par biolistique les EPRV dans les cellules d'une plante saine dont le génome ne contient pas les EPRV introduits. Une expérience similaire a été réalisée à partir d'un EPRV du PVCV, et a conduit à une infection systémique par le virus (Richert-Pöggeler et al., 2003). L'EPRV BSImV récemment séquencé a été amplifié par long-range PCR puis cloné dans un vecteur plasmidique. La construction a été ensuite bombardée sur 38 jeunes plants *M. acuminata*. Les résultats préliminaires semblent prometteurs, car en faveur d'une initiation d'infection systémique par le BSImV, détectée par IC-PCR sur broyat de feuilles et PCR sur ADN totaux. Les analyses complémentaires concernant la détection du génome viral complet, des transcrits et des protéines virales ainsi que la visualisation des particules virales par microscopie électronique sont en cours.

La structure de l'EPRV BSImV s'est révélée être plus simple que celle de l'EPRV BSGFV. En effet, cet EPRV est principalement constitué d'un seul fragment correspondant à une fois et demi le génome viral. L'activation de cet EPRV pourrait donc ne pas resulter d'évènements de recombinaison homologue. Une transcription directe de cet EPRV semble être l'hypothèse la plus simple qu'il faudra tester par la

recherche de transcrits de l'EPRV dans différents tissus du bananier cv. PKW, ainsi que chez les hybrides interspécifiques.

Mécanismes de défense des bananiers contre les EPRV infectieux

A long terme :

Une meilleure connaissance de ces mécanismes sera en effet utile pour comprendre comment ont évolué les EPRV infectieux et quelles sont les pressions de sélection qui leurs sont associées.

Un projet mise en place dans l'équipe 1 où s'est effectué ce travail de thèse, concerne l'origine moléculaire et génétique du contrôle des EPRV infectieux par les bananiers. Ce contrôle est intimement lié à la résistance des *M. balbisiana* aux BSV issus de l'activation des EPRV, ou d'une transmission par cochenilles, ainsi qu'aux différents facteurs responsables de l'activation des EPRV infectieux. Afin d'explorer cet axe de recherche, un chercheur spécialiste en expression et régulation génétique a été récemment recruté dans l'équipe.

Plusieurs pistes peuvent être explorées, si l'on considère que des mécanismes génétiques ou épigénétiques sont en jeu. Une première étude sur les mécanismes génétiques de résistance consisterait à cloner le facteur BEL, mis en évidence dernièrement dans l'équipe (Lheureux et al., 2003). Sa présence sous forme hétérozygote chez le parent *M. balbisiana* confèrerait une résistance à l'activation des EPRV, tandis qu'un des allèles sous forme homozygote dans la descendance permettrait l'activation des EPRV infectieux et la multiplication virale des espèces BSOLV, BSGFV et BSImV. La deuxième piste concerne l'identification de mécanismes épigénétiques associés à la régulation des EPRV infectieux chez le cv. PKW, comme une méthylation des séquences intégrées, l'association des EPRV à des histones liées au silencing de l'expression des gènes, la transcription des EPRV et/ou la production de siARN à partir de ces transcrits EPRV, comme ce qui est observé pour les EPRV des solanacées (Mette et al., 2002; Noreen et al., 2007; Richert-Pöggeler et al., 2003; Staginnus et al., 2007). Il est également possible que des EPRV infectieux différents soient régulés par des mécanismes moléculaires distincts en lien avec leur structure et leur localisation dans le génome.

Biologie évolutive des EPRV infectieux

A long terme :

Nos connaissances sur la biologie des EPRV ont largement progressé ces dernières années. La présence d'EPRV infectieux dans le génome des plantes soulève de nombreuses questions évolutives liées notamment au coût de la résistance envers les EPRV infectieux. Un point essentiel pour aborder ces questions concerne l'étude de l'impact des EPRV infectieux en conditions naturelles. En effet, la culture des bananiers s'éloigne beaucoup des conditions qui ont vu l'émergence des intégrations. Les bananiers cultivés sont premièrement domestiqués, ce qui a pu entraîner des changements chromosomiques, des balayages sélectifs, des épisodes de dérive génétique propices à la perte ou à la fixation des EPRV infectieux en relativement peu de temps. De plus, la domestication est le fruit de nombreux croisements entre sous-espèces, et entre espèces, ce qui rend très difficile le suivi *a posteriori* des filiations et la reconstruction phylogénétique utiles à l'interprétation de la distribution, perte ou éventuellement amplification des EPRV attendus chez les bananiers. Enfin, les bananiers cultivés sont stériles et propagés clonalement par multiplication végétative, alors que les bananiers sauvages ont une reproduction sexuée en plus de la multiplication végétative.

Il serait dès lors intéressant d'étudier la prévalence du BSV dans les bananeraies sauvages, en populations naturelles de bananier, et de déterminer le coût associé aux infections, chez *M. acuminata*. Ensuite, une recherche de la fréquence des hybridations interspécifiques naturelles pourrait être conduite dans les zones de sympatries entre *M. balbisiana* et *M. acuminata* ssp. *burmanica / burmanicoides* (nord de l'Inde), avec *M. acuminata* ssp. *siamea* (sud de la Chine), *M. acuminata* ssp. errans (Philippines) et *M. acuminata* ssp. *banksii* (Papouasie Nouvelle Guinée). Nous pourrons ensuite vérifier que les EPRV infectieux des hybrides naturels sont activés et provoquent bien une infection systémique du BSV.

Enfin, bien que les badnavirus soient un genre viral émergent, que le BSV affecte une culture économiquement importante, et que les EPRV infectieux constituent une singularité dans le monde viral, la biologie générale du BSV reste à ce jour très peu connue. Des études du cycle de multiplication, de la localisation tissulaire et

cellulaire du BSV lors de l'infection, ou de génomique (fonction des protéines virales, présence de protéine suppresseur de silencing) sont en effet nécessaires à la compréhension de la complexité de ce modèle biologique, quels que soient la discipline, les questions de recherche et les champs d'application.

Inventaire de la biodiversité virale intégrée : étude de la nature et des structures des EPRV

A court terme :

Le séquençage des clones BAC contenant les EPRV d'autres espèces virales recherchées : BSOLV, BSImV, BSMyV entre dans sa dernière phase, ce qui laisse entrevoir à courts terme un élargissement de nos connaissances sur la biodiversité des intégrations du BSV chez *M. balbisiana*. Les résultats préliminaires d'analyse des Fingerprint et de séquençage des clones BAC contenant les EPRV des autres espèces (projet en collaboration avec l'UMR DAP au Cirad) semblent également montrer un faible nombre de locus d'intégration (1 voire 2) pour chaque espèce virale, comme ce que nous avons montré pour l'EPRV BSGFV. C'est le cas par exemple des intégrations BSImV. Pour cette espèce, les fingerprint sur les 24 BAC contenant des séquences de BSImV ont tous un profil identique. Un clone BAC a pu être séquencé par le Génoscope, et annoté avec l'aide de Franck-Christophe Baurens et Stéphanie Sidibe-Bocs à l'UMR DAP. La séquence de l'EPRV a été analysée lors du stage de Master 2 d'Olivier Guidolin, que j'ai co-encadré. De manière analogue aux EPRV BSGFV, les résultats préliminaires de ségrégation génétique des EPRV de l'espèce BSImV suggèrent la présence d'un seul locus homozygote chez *M. balbisiana* cv. PKW. Les données de séquences des autres espèces de BSV intégrées chez le cv. PKW sont en cours de production, et devraient fournir une vue d'ensemble des intégrations chez ce cultivar et permettre de déterminer les caractéristiques communes de structure, de sites d'intégration et de nombre d'EPRV par génome. Il sera également intéressant de connaître plus précisément les lieux d'intégration au sein du génome, afin de vérifier si les intégrations BSV sont, à l'instar des EPRV des solanacées, majoritairement localisées dans l'hétérochromatine ou dans des régions riches en rétroéléments. Cette localisation pourra notamment être précisée par des

analyses de type FISH, en hybridant directement les clones BAC sur les chromosomes du cv. PKW, ou mieux, en hybridant les EPRV seulement, obtenus après sous-clonage. Certains clones BAC portant les EPRV sont en effet riches en éléments transposables ce qui pourrait induire un bruit de fond important par des hybridations non désirées dans le génome.

A court terme :

Afin de rechercher la présence d'un EPRV d'une espèce connue dans un génome inconnu, une alternative aux fingerprint et à l'utilisation de sondes radioactives pourrait être la Q-PCR. Cette méthode peut être utilisée pour quantifier de manière absolue le nombre de copies par génome d'un gène lorsque la séquence du fragment-cible est connue (Chiang et al., 1996; Ingham et al., 2001). De mise en place relativement longue (mise au point des couples d'amorces spécifiques des différentes espèces de BSV intégré, et du standard interne), cette méthode permet de fournir rapidement des résultats sur un grand nombre d'individus. Il sera donc possible de comparer pour un EPRV donné, la quantité de copies existantes entre les génotypes de chaque espèce de bananier, et de détecter une amplification de l'EPRV dans certains génomes, comme ce qui est observé chez les *Ns*EPRV et *Nto*EPRV.

A court terme :

L'approche classique de détection de la diversité des EPRV par PCR dégénérées sur des bananiers sains a récemment fait ses preuves, puisque des séquences appartenant à des dizaines de nouvelles espèces potentielles de BSV ont été mises à jour à partir d'un nombre très limité d'espèces de bananiers (Geering et al., 2005a). Il serait utile de conduire un inventaire exhaustif de la biodiversité intégrée par le criblage du plus grand nombre d'espèces des genres *Musa* et *Ensete* disponibles dans les collections mondiales.

A plus long terme :

Les recherches sur les EPRV se sont principalement focalisées sur la caractérisation des intégrations infectieuses présentes uniquement à ce jour dans le génome B. Or, le génome de *M. acuminata* contient également de nombreux EPRV, et mérite une

attention particulière pour deux raisons. La première vient du fait que les collectes d'isolats viraux et les études phylogénétiques menées ces dernières années sur le BSV ont révélé la présence dans le génome A d'EPRV correspondant à environ une dizaine d'espèces jusqu'alors inconnues. Il sera utile de vérifier si ces EPRV correspondent à des virus existants encore aujourd'hui. Pour cela, il faudra rechercher leur présence dans des épidémies de BSV. Ensuite, nous pourrons démontrer si certains de ces EPRV sont infectieux après CIV de différents génotypes contenant du génome A, en lien avec leurs strutures et leur capacité à reconstituer un génome viral dans sa totalité. La CIV constitue en effet un stress relativement simple à mettre en œuvre, qui amplifie efficacement le phénomène d'activation des EPRV (Dallot et al., 2001). Si l'on transpose au génome A l'hypothèse que l'activation des EPRV du génome B est liée au stress génomique provoqué par des hybridations interspécifiques, il conviendra de privilégier lors du test d'activation les hybrides intra- et interspécifiques, dits 'naturels' (*i.e.* anciens) mais également récemment créés par l'homme. Enfin, les génotypes devront être choisis pour maximiser la diversité du génome A afin de ne pas passer à côté d'EPRV infectieux ; et donc provenir des neuf sous-espèces décrites de l'espèce *M. acuminata.*

La deuxième raison qui doit nous inciter à étudier les EPRV du génome A est en lien étroit avec l'avancée du projet de séquençage du génome de la banane. D'ici environ 2010, le génome du cultivar haploïde doublé 'Pahang' (*M. acuminata* ssp. *malaccencis*) sera séquencé. Une approche bioinformatique sera alors nécessaire pour détecter les séquences pararétrovirales intégrées au génome, puis pour les annoter, et enfin pour analyser et comparer leurs caractéristiques de manière précise et exhaustive. Il sera alors possible de connaitre le nombre d'EPRV, la nature des espèces intégrées, le patron de dispersion dans le génome (en clusters ou aléatoire), et la présence de sites d'intégrations privilégiés (par exemple la richesse en GC, la présence de poly-AT, ou une intégration proche d'un rétro-élément), mais également leur taille (génome complet ou tronqué) et leur structure (fragments réarrangés ou non). Tous ces éléments permettront d'avoir une vision globale des intégrations utile pour comprendre les mécanismes liés aux EPRV. D'un point de vue appliqué, ces données seront utiles pour détecter les EPRV potentiellement infectieux. A plus long terme, la

séquence du génome A pourra servir de base au séquençage de l'espèce *M. balbisiana*. Il reste en effet de nombreuses espèces virales à découvrir dans le génome de cette dernière espèce. Dans l'attente de son séquençage, il serait nécessaire de caractériser les EPRV des autres espèces, en hybridant la banque BAC de ce cultivar avec des séquences virales correspondant aux espèces récemment découvertes. Cette étude nous permettra de connaitre rapidement le nombre de locus d'intégration et les espèces intégrées.

Application : Utilisation des EPRV comme marqueur de phylogénie

A plus long terme :

Une fois la biodiversité des espèces intégrées mieux connue, il sera envisageable d'utiliser les EPRV comme marqueurs de phylogénie des bananiers. Les intégrations devront être suffisamment anciennes et partagées par un maximun d'espèces. Les EPRV choisis devront donc être intégrés avant la spéciation *M. acuminata*/*M. balbisiana*, voire partagés entre différentes espèces de bananiers, ou même être communs au genre *Ensete* et *Musa*. Nous pourrons alors tester si ces EPRV sont de bons marqueurs de phylogénie du genre, suffisamment polymorphes et informatifs pour résoudre les ordres d'embranchements des espèces qui sont encore approximatifs à l'heure actuelle.

REFERENCES BIBLIOGRAPHIQUES

Almeida, R. & Allshire, R. C. (2005). RNA silencing and genome regulation. *Trends Cell Biol* **15**, 251-258.

Amedeo, P., Habu, Y., Afsar, K., Scheid, O. M. & Paszkowski, J. (2000). Disruption of the plant gene MOM releases transcriptional silencing of methylated genes. *Nature* **405**, 203-206.

Annaheim, M. & Lanzrein, B. (2007). Genome organization of the *Chelonus inanitus polydnavirus*: excision sites, spacers and abundance of proviral and excised segments. *J Gen Virol* **88**, 450-457.

Argent, G. C. G. (1976). The wild bananas of Papua New Guinea. *Notes Roy Bot Gard Edinburgh* **35**, 77 - 114.

Ashby, M. K., Warry, A., Bejarano, E. R., Khashoggi, A., Burrell, M. & Lichtenstein, C. P. (1997). Analysis of multiple copies of geminiviral DNA in the genome of four closely related *Nicotiana* species suggest a unique integration event. *Plant Mol Biol* **35**, 313-321.

Astier, S., Albouy, J., Maury, Y. & Lecoq, H. (2001). *Principes de virologie végétale: génome, pouvoir pathogène, écologie des virus*: Inra Editions.

Banks, D. J., Beres, S. B. & Musser, J. M. (2002). The fundamental contribution of phages to GAS evolution, genome diversification and strain emergence. *Trends Microb* **10**, 515-521.

Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., Romero, D. A. & Horvath, P. (2007). CRISPR provides acquired resistance against viruses in prokaryotes. *Science* **315**, 1709-1712.

Baulcombe, D. (2004). RNA silencing in plants. *Nature* **431**, 356-363.

Bejarano, E. R., Khashoggi, A., Witty, M. & Lichtenstein, C. (1996). Integration of multiple repeats of geminiviral DNA into the nuclear genome of tobacco during evolution. *Proc Natl Acad Sci U S A* **93**, 759-764.

Belshaw, R., Pereira, V., Katzourakis, A., Talbot, G., Paces, J., Burt, A. & Tristem, M. (2004). Long-term reinfection of the human genome by endogenous retroviruses. *Proc Natl Acad Sci U S A* **101**, 4894-4899.

Bennetzen, J. L. (2000). Transposable element contributions to plant gene and genome evolution. *Plant Mol Biol* **42**, 251-269.

Bergh, O., Borsheim, K. Y., Bratbak, G. & Heldal, M. (1989). High abundance of viruses found in aquatic environments. *Nature* **340**, 467-468.

Best, S., Le Tissier, P. R. & Stoye, J. P. (1997). Endogenous retroviruses and the evolution of resistance to retroviral infection. *Trends Microb* **5**, 313-318.

Blaise, S., de Parseval, N., Benit, L. & Heidmann, T. (2003). Genomewide screening for fusogenic human endogenous retrovirus envelopes identifies syncytin 2, a gene conserved on primate evolution. *Proc Natl Acad Sci U S A* **100**, 13013-13018.

Blond, J.-L., Lavillette, D., Cheynet, V., Bouton, O., Oriol, G., Chapel-Fernandes, S., Mandrand, B., Mallet, F. & Cosset, F.-L. (2000). An envelope glycoprotein of the Human Endogenous Retrovirus HERV-W is expressed in the human placenta and fuses cells expressing the type D mammalian retrovirus receptor. *Journal of virology* **74**, 3321-3329.

Bouhida, M., Lockhart, B. E. L. & Olszewski, N. E. (1993). An analysis of the complete sequence of a *Sugarcane bacilliform virus* genome infectious to banana and rice. *J Gen Virol* **74**, 15-22.

Bousalem, M., Douzery, E. J. P. & Seal, S. E. (2008). Taxonomy, molecular phylogeny and evolution of plant reverse transcribing viruses (family *Caulimoviridae*) inferred from full-length genome and reverse transcriptase sequences. *Arch Virol* **153**, 1085-1102.

Bromham, L. (2002). The human zoo: endogenous retroviruses in the human genome. *Trends Ecol Evol* **17**, 91-97.

Budiman, M. A., Mao, L., Wood, T. C. & Wing, R. A. (2000). A deep-coverage tomato BAC library and prospects toward development of an STC framework for genome sequencing. *Genome Res* **10**, 129-136.

Campbell, A. (2003). The future of bacteriophage biology. *Nat Rev Genet* **4**, 471-477.

Capy, P., Gasperi, G., Biemont, C. & Bazin, C. (2000). Stress and transposable elements: co-evolution or useful parasites? *Heredity* **85**, 101-106.

Carreel, F. (1994). Etude de la diversité génétique des bananiers (genre *Musa*) à l'aide des marqueurs RFLP. PhD Thesis: Institut National Agronomique Paris-Grignon.

Cheesman, E. E. (1947). Classification of the banana. *Kew bulletin* **2**, 97-117.

Cheng, C. P., Lockhart, B. E. & Olszewski, N. E. (1996). The ORF I and II proteins of *Commelina yellow mottle virus* are virion-associated. *Virology* **223**, 263-271.

Chiang, P. W., Song, W. J., Wu, K. Y., Korenberg, J. R., Fogel, E. J., Van Keuren, M. L., Lashkari, D. & Kurnit, D. M. (1996). Use of a fluorescent-PCR reaction to detect genomic sequence copy number and transcriptional abundance. *Genome Res* **6**, 1013-1026.

Claverie, J. M. (2006). Viruses take center stage in cellular evolution. *Genome Biol* **7**, 110.

Coustau, C., Chevillon, C. & Ffrench-Constant, R. (2000). Resistance to xenobiotics and parasites: can we count the cost? *Trends Ecol Evol* **15**, 378-383.

Cox-Foster, D. L., Conlan, S., Holmes, E. C., Palacios, G., Evans, J. D., Moran, N. A., Quan, P. L., Briese, T., Hornig, M., Geiser, D. M., Martinson, V., vanEngelsdorp, D., Kalkstein, A. L., Drysdale, A., Hui, J., Zhai, J. H., Cui, L. W., Hutchison, S. K., Simons, J. F., Egholm, M., Pettis, J. S. & Lipkin, W. I. (2007). A metagenomic survey of microbes in honey bee colony collapse disorder. *Science* **318**, 283-287.

Crochu, S., Cook, S., Attoui, H., Charrel, R. N., De Chesse, R., Belhouchet, M., Lemasson, J.-J., de Micco, P. & de Lamballerie, X. (2004). Sequences of flavivirus-related RNA viruses persist in DNA form integrated in the genome of *Aedes* spp. mosquitoes. *J Gen Virol* **85**, 1971-1980.

Dahal, G., Hughes, d. A., Thottappilly, G. & Lockhart, B. E. L. (1998). Effect of temperature on symtom expression and expression and reliability of *Banana streak badnavirus* detection in naturally infected plantain and banana (*Musa* spp). *Plant Disease* **82**, 16-21.

Dallot, S., Acuna, P., Rivera, C., Ramirez, P., Cote, F., Lockhart, B. E. L. & Caruana, M. L. (2001). Evidence that the proliferation stage of micropropagation procedure is determinant in the expression of *Banana streak virus* integrated into the genome of the FHIA 21 hybrid (*Musa* AAAB). *Arch Virol* **146**, 2179-2190.

Daniells, J., Thomas, J. E. & Smith, M. (1995). Seed transmission of *Banana streak virus* confirmed. *Infomusa* **4**, 7.

Daniells, J. W., Geering, A. D. W., Bryde, N. J. & Thomas, J. E. (2001). The effect of *Banana streak virus* on the growth and yield of dessert bananas in tropical Australia. *Ann Appl Biol* **139**, 51-60.

Daubin, V. & Ochman, H. (2004). Start-up entities in the origin of new genes. *Current opinion in genetics & development* **14**, 616-619.

de Parseval, N. & Heidmann, T. (2005). Human endogenous retroviruses: from infectious elements to human genes. *Cytogenet Genome Res* **110**, 318-332.

Delanoy, M., Salmon, M., Kummert, J., Frison, E. & Lepoivre, P. (2003). Development of real-time PCR for the rapid detection of episomal *Banana streak virus* (BSV). *Plant Disease* **87**, 33-38.

Denham, T. P., Haberle, S. G., Lentfer, C., Fullagar, R., Field, J., Therin, M., Porch, N. & Winsborough, B. (2003). Origins of agriculture at Kuk Swamp in the highlands of New Guinea. *Ann Bot (Lond)* **301**, 189-193.

Di Franco, C., Pisano, C., Fourcade-Peronnet, F., Echalier, G. & Junakovic, N. (1992). Evidence for de novo rearrangements of Drosophila transposable elements induced by the passage to the cell culture. *Genetica* **87**, 65-73.

Douchkov, D., Nowara, D., Zierold, U. & Schweizer, P. (2005). A High-Throughput Gene-Silencing System for the Functional Assessment of Defense-Related Genes in Barley Epidermal Cells, pp. 755-761.

Drezen, J. M., Provost, B., Espagne, E., Cattolico, L., Dupuy, C., Poirie, M., Periquet, G. & Huguet, E. (2003). Polydnavirus genome: integrated vs. free virus. *Journal of insect physiology* **49**, 407-417.

Duffy, S., Shackelton, L. A. & Holmes, E. C. (2008). Rates of evolutionary change in viruses: patterns and determinants. *Nat Rev Genet* **9**, 267-276.

Dupuy, C., Huguet, E. & Drezen, J.-M. (2006). Unfolding the evolutionary story of polydnaviruses. *Virus Res* **117**, 81-89.

Edwards, R. A. & Rohwer, F. (2005). Viral metagenomics. *Nat Rev Micro* **3**, 504-510.

Espagne, E., Dupuy, C., Huguet, E., Cattolico, L., Provost, B., Martins, N., Poirie, M., Periquet, G. & Drezen, J. M. (2004). Genome sequence of a polydnavirus: insights into symbiotic virus evolution. *Science* **306**, 286-289.

Fargette, D., Konate, G., Fauquet, C., Muller, E., Peterschmitt, M. & Thresh, J. M. (2006). Molecular ecology and emergence of tropical plant viruses. *Annu Rev Phytopathol* **44**, 235-260.

Fauquet, C. M., Mayo, M. A., Maniloff, J., Desselberger, U. & Ball, L. A. (2005). Virus taxonomy : eighth report of the international committee on taxonomy viruses. *Elsevier, Academic Press.*

Federici, B. A. & Bigot, Y. (2003). Origin and evolution of polydnaviruses by symbiogenesis of insect DNA viruses in endoparasitic wasps. *J Insect Physiol* **49**, 419-432.

Flint, S. J., Enquist, L. W. & Skalka, A. M. (2003). *Principles of virology: molecular biology, pathogenesis, and control of animal viruses, 2nd Edition*: ASM Press.

Folliot, M., Galzi, S., Laboureau, N., Caruana, M.-L., Teycheney, P.-Y. & Côte, F.-X. (2005). Risk assessment of spreading *Banana streak virus* (BSV) through in vitro culture. July 23-28. In *XIIIth International Congress of Virology*. San Fransisco (USA).

Forterre, P. (2006). The origin of viruses and their possible roles in major evolutionary transitions. *Virus Res* **117**, 5-16.

Gallei, A., Pankraz, A., Thiel, H. J. & Becher, P. (2004). RNA recombination in vivo in the absence of viral replication. *Journal of virology* **78**, 6271-6281.

Gawel, N. J., Jarret, R. L. & Whittemore, A. P. (1992). Restriction fragment length polymorphism (RFLP)-based phylogenetic analysis of *Musa*. *Theor Appl Genet* **84**, 286-290.

Gayral, P., Noa-Carrazana, J.-C., Lescot, M., Lheureux, F., Lockhart, B. E. L., Matsumoto, T., Piffanelli, P. & Iskra-Caruana, M.-L. (2008). A single *Banana streak virus* integration event in the banana genome as the origin of infectious endogenous pararetrovirus. *Journal of virology* **82**, 6697-6710.

Geering, A. D. W., McMichael, L. A., Dietzgen, R. G. & Thomas, J. E. (2000). Genetic diversity among Banana streak virus isolates from Australia. *Phytopathology* **90**, 921-927.

Geering, A. D. W., Olszewski, N. E., Dahal, G., Thomas, J. E. & Lockhart, B. E. L. (2001). Analysis of the distribution and structure of integrated *Banana streak virus* DNA in a range of *Musa* cultivars. *Mol Plant Pathol* **2**, 207-213.

Geering, A. D. W., Olszewski, N. E., Harper, G., Lockhart, B. E. L., Hull, R. & Thomas, J. E. (2005a). Banana contains a diverse array of endogenous badnaviruses. *J Gen Virol* **86**, 511-520.

Geering, A. D. W., Pooggin, M. M., Olszewski, N. E., Lockhart, B. E. L. & Thomas, J. E. (2005b). Characterisation of *Banana streak Mysore virus* and evidence that its DNA is integrated in the B genome of cultivated *Musa*. *Arch Virol* **150**, 787-796.

Ghosh, P., Wasil, L. R. & Hatfull, G. F. (2006). Control of phage Bxb1 excision by a novel recombination directionality factor. *Plos Biol* **4**, 964-974.

Gowen, S. (1995). *Bananas and plantains, international network for improvement of banana and plantain.* London ; New York: Chapman & Hall.

Grandbastien, M.-A. (1998). Activation of plant retrotransposons under stress conditions. *Trends Plant Sci* **3**, 181-187.

Grandbastien, M. A., Audeon, C., Bonnivard, E., Casacuberta, J. M., Chalhoub, B., Costa, A. P., Le, Q. H., Melayah, D., Petit, M., Poncet, C., Tam, S. M., Van Sluys, M. A. & Mhiri, C. (2005). Stress activation and genomic impact of Tnt1 retrotransposons in Solanaceae. *Cytogenet Genome Res* **110**, 229-241.

Gregor, W., Mette, M. F., Staginnus, C., Matzke, M. A. & Matzke, A. J. M. (2004). A distinct endogenous pararetrovirus family in *Nicotiana tomentosiformis*, a diploid progenitor of polyploid tobacco. *Plant Physiol* **134**, 1191-1199.

Griffiths, D. J. (2001). Endogenous retroviruses in the human genome sequence. *Genome Biol* **2**, REVIEWS1017.

Gruber, A., Stettler, P., Heiniger, P., Schumperli, D. & Lanzrein, B. (1996). Polydnavirus DNA of the braconid wasp *Chelonus inanitus* is integrated in the wasp's genome and excised only in later pupal and adult stages of the female. *J Gen Virol* **77 (Pt 11)**, 2873-2879.

Hagen, L. S., Jacquemond, M., Lepingle, A., Lot, H. & Tepfer, M. (1993). Nucleotide sequence and genomic organization of *Cacao swollen shoot virus*. *Virology* **196**, 619-628.

Hansen, C. & Heslop-Harrison, J. S. (2004). Sequences and phylogenies of plant pararetroviruses, viruses, and transposable elements. In *Advances in Botanical Research Incorporating Advances in Plant Pathology, Vol 41*, pp. 165-193.

Hansen, C. N., Harper, G. & Heslop-Harrison, J. S. (2005). Characterisation of pararetrovirus-like sequences in the genome of potato (*Solanum tuberosum*). *Cytogenet Genome Res* **110**, 559-565.

Harper, G., Dahal, G., Thottappilly, G. & Hull, R. (1999a). Detection of episomal *Banana streak badnavirus* by IC-PCR. *J Gen Virol* **79**, 1-8.

Harper, G., Hart, D., Moult, S. & Hull, R. (2002). Detection of *Banana streak virus* in field samples of bananas from Uganda. *Ann Appl Biol* **141**, 247-257.

Harper, G., Hart, D., Moult, S. & Hull, R. (2004). *Banana streak virus* is very diverse in Uganda. *Virus Res* **100**, 51-56.

Harper, G., Hart, D., Moult, S., Hull, R., Geering, A. & Thomas, J. (2005). The diversity of *Banana streak virus* isolates in Uganda. *Arch Virol*, -46.

Harper, G. & Hull, R. (1998). Cloning and sequence analysis of banana streak virus DNA. *Virus Genes* **17**, 271-278.

Harper, G., Osuji, J. O., Heslop-Harrison, J. S. P. & Hull, R. (1999b). Integration of *Banana streak badnavirus* into the *Musa* genome: molecular and cytogenetic evidence. *Virology* **255**, 207-213.

Heslop-Harrison, J. S. & Schwarzacher, T. (2007). Domestication, genomics and the future for banana. *Ann Bot (Lond)*.

Hohn, T., Richert-Pöggeler, K. R., Harper, G., Schwarzacher, T., Teo, C. H., Techeney, P. Y., Iskra-Caruana, M. L. & Hull, R. (2008). Evolution of integrated plant viruses. In Virus Evolution. Springer, Heidelberg, M. Roosinck ed.

Horrocks, M., Bulrner, S. & Gardner, R. O. (2008). Plant microfossils in prehistoric archaeological deposits from Yuku rock shelter, Western Highlands, Papua New Guinea. *J Archeol Sci* **35**, 290-301.

Horry, J. P. & Jay, M. (1997). An evolutionary background for bananas as deduced from flavonoids diversification. In: Jarret, R.L. (Ed). Identification of genetic diversity in the genus *Musa*, Proceedings of an international workshop held at Los Banos, Philippines, 5-10 september. INIBAP, Montferrier sur Lez, France, pp. 41-55.

Huang, Q. & Hartung, J. S. (2001). Cloning and sequence analysis of an infectious clone of *Citrus yellow mosaic virus* that can infect sweet orange via *Agrobacterium*-mediated inoculation. *J Gen Virol* **82**, 2549-2558.

Hull, R. (2002). *Matthew's plant virology, 4th edition*: San Diego, San Francisco, New York, Boston, London, Sydney, Tokyo : Academic Press.

Hull, R., Harper, G. & Lockhart, B. (2000). Viral sequences integrated into plant genomes. *Trends Plant Sci* **5**, 362-365.

Ingham, D. J., Beer, S., Money, S. & Hansen, G. (2001). Quantitative real-time PCR assay for determining transgene copy number in transformed plants. *Biotechniques* **31**, 132-+.

Iskra-Caruana, M. L., Lheureux, F. & Teycheney, P. Y. (2003). Les pararétrovirus endogènes (EPRV), voie nouvelle de transmission des virus de plantes. *Virologie* **7**.

Iskra Caruana M.L., Lheureux F., Noa-Carrazana J.C., Piffanelli P., Carreel F., Jenny C., Laboureau N. & B.E.L., L. (2003). Unstable balance of relation between parretrovirus and its host plant : the BSV-EPRV banana pathosystem : [Abstract], May 28-31. *EMBO Workshop Genomic Approaches in Plant Virology.*

Jacquot, E., Dautel, S., Leh, V., Geldreich, A., Yot, P. & Keller, M. (1997). Les pararétrovirus de plantes. *Virologie* **1**, 111-120.

Jacquot, E., Hagen, L. S., Jacquemond, M. & Yot, P. (1996). The open reading frame 2 product of *Cacao swollen shoot badnavirus* is a nucleic acid-binding protein. *Virology* **225**, 191-195.

Jakowitsch, J., Mette, M. F., van der Winden, J., Matzke, M. A. & Matzke, A. J. M. (1999). Integrated pararetroviral sequences define a unique class of dispersed repetitive DNA in plants. *Proc Natl Acad Sci U S A* **96**, 13241-13246.

Jaufeerally-Fakim, Y., Khorugdharry, A. & Harper, G. (2006). Genetic variants of *Banana streak virus* in Mauritius. *Virus Res* **115**, 91-98.

Johnson, W. E. & Coffin, J. M. (1999). Constructing primate phylogenies from ancient retrovirus sequences. *Proc Natl Acad Sci U S A* **96**, 10254-10260.

Jones, D. R. (2000). *Diseases of banana, abacá, and enset.* Wallingford, Oxon, UK ; New York: CABI Pub.

Kaku, H., Nishizawa, Y., Ishii-Minami, N., Akimoto-Tomiyama, C., Dohmae, N., Takio, K., Minami, E. & Shibuya, N. (2006). Plant cells recognize chitin fragments for defense signaling through a plasma membrane receptor, pp. 11086-11091.

Katzourakis, A., Rambaut, A. & Pybus, O. G. (2005). The evolutionary dynamics of endogenous retroviruses. *Trends Microb* **13**, 463-468.

Keeling, P. J. & Slamovits, C. H. (2005). Causes and effects of nuclear genome reduction. *Current opinion in genetics & development* **15**, 601-608.

Kenton, A., Khashoggi, A., Parokonny, A., Bennett, M. D. & Lichtenstein, C. (1995). Chromosomal location of endogenous geminivirus-related DNA sequences in *Nicotiana tabacum* L. *Chromosome Res* **3**, 346-350.

Kenyon, L., Lebas, B. S. & Seal, S. E. (2008). Yams (*Dioscorea* spp.) from the South Pacific Islands contain many novel badnaviruses: implications for international movement of yam germplasm. *Arch Virol* **153**, 877-889.

Kidwell, M. G. & Lisch, D. R. (2000). Transposable elements and host genome evolution. *Trends Ecol Evol* **15**, 95-99.

Kim, F. J., Battini, J. L., Manel, N. & Sitbon, M. (2004). Emergence of vertebrate retroviruses and envelope capture. *Virology* **318**, 183-191.

Koonin, E. V. & Dolja, V. V. (2006). Evolution of complexity in the viral world: the dawn of a new vision. *Virus Res* **117**, 1-4.

Koonin, E. V. & Martin, W. (2005). On the origin of genomes and cells within inorganic compartments. *Trends Genet* **21**, 647-654.

Kubiriba, J., Legg, J. P., Tushemereirwe, W. & Adipala, E. (2001). Vector transmission of *Banana streak virus* in the screenhouse in Uganda. *Ann Appl Biol* **139**, 37-43.

Kunii, M., Kanda, M., Nagano, H., Uyeda, I., Kishima, Y. & Sano, Y. (2004). Reconstruction of putative DNA virus from endogenous rice tungro bacilliform virus-like sequences in the rice genome: implications for integration and evolution. *BMC Genomics* **5**, 80.

Lafleur, D. A., Lockhart, B. E. L. & Olszewski, N. E. (1996). Portions of *Banana streak badnavirus* genome are integrated in the genome of its host *Musa* sp. *Phytopathology (supplement)* **86**, 100.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W. et al. (2001). Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921.

Le Provost, G., Iskra-Caruana, M. L., Acina, I. & Teycheney, P. Y. (2006). Improved detection of episomal Banana streak viruses by multiplex immunocapture PCR. *J Virol Methods* **137**, 7-13.

Leitch, A. R. (2007). Conserved gene order belies rapid genome turnover - The dynamic interplay between genomic DNA and the outside world. *Heredity* **98**, 61-62.

Lentfer, C. J. & Green, R. C. (2004). Phytoliths and the evidence for banana cultivation at the Lapita Reber-Rakival site on Watom Island, Papua New Guinea. *Rec Aust Mus*, 75-88.

Lescot, M., Piffanelli, P., Ciampi, A. Y., Ruiz, M., Blanc, G., Leebens-Mack, J., da Silva, F. R., Santos, C. M., D'Hont, A., Garsmeur, O., Vilarinhos, A. D., Kanamori, H., Matsumoto, T., Ronning, C. M., Cheung, F., Haas, B. J., Althoff, R., Arbogast, T., Hine, E., Pappas, G. J., Sasaki, T., Souza, M. T., Miller, R. N., Glaszmann, J. C. & Town, C. D. (2008). Insights into the *Musa* genome: syntenic relationships to rice and between *Musa* species. *BMC Genomics* **9**, 58.

Lewinski, M. K., Yamashita, M., Emerman, M., Ciuffi, A., Marshall, H., Crawford, G., Collins, F., Shinn, P., Leipzig, J., Hannenhalli, S., Berry, C. C., Ecker, J. R. & Bushman, F. D. (2006). Retroviral DNA integration: viral and cellular determinants of target-site selection. *Plos Pathogens* **2**, e60.

Lheureux, F. (2002). Etude des mécanismes génétiques impliqués dans l'expression des séquences EPRVs pathogènes des Bananiers au cours de croisements génétiques interspécifiques. In *Ecole Nationale Supérieure Agronomique de Momntpellier*, p. 102. Montpellier: Université Sciences et Techniques du Languedoc USTL.

Lheureux, F., Carreel, F., Jenny, C., Lockhart, B. & Iskra-Caruana, M. (2003). Identification of genetic markers linked to banana streak disease expression in inter-specific *Musa* hybrids. *Theor Appl Genet* **106**, 594-598.

Lheureux, F., Laboureau, N., Muller, E., Lockhart, B. E. & Iskra-Caruana, M. L. (2007). Molecular characterization of *Banana streak acuminata Vietnam virus* isolated from *Musa acuminata siamea* (banana cultivar). *Arch Virol* **152**, 1409-1416.

Lockhart, B. E., Menke, J., Dahal, G. & Olszewski, N. E. (2000). Characterization and genomic analysis of *Tobacco vein clearing virus,* a plant pararetrovirus that is transmitted vertically and related to sequences integrated in the host genome. *J Gen Virol* **81**, 1579-1585.

Lockhart, B. E. L. (1990). Evidence for a double-stranded circular DNA Genome in a second Group of plant Viruses. *Phytopathology* **80**, 127-131.

Lockhart, B. E. L. (1994). *Banana streak virus*. In: Compendium of tropical plant diseases: R. C. Ploetz, G. A. Gentmyer, W. T. Nishijima, K. G. Rohrbach H. D. Ohr, APS Press, St Paul.

Lockhart, B. E. L. (1995). Banana streak badnavirus infection in *Musa*: epidemiology, diagnosis and control. *ASPAC Food and fertilizerr technology center (taiwan) technical bulletin* **143**, 1-11.

Lockhart, B. E. L. & Jones, D. R. (2000a). Diseased caused by virus: banana mosaic. In *Diseases of banana, abacà and enset*, pp. 256-263. Edited by D. R. Jones. New york: CABI Publishing.

Lockhart, B. E. L. & Jones, D. R. (2000b). Diseases caused by virus: banana streak. In *Diseases of banana, abaca and enset*. Edited by D. R. Jones. New York: CABI Publishing.

Lockhart, B. E. L. & Lesemann, D. E. (1998). Occurrence of *Petunia vein-clearing virus* in the U.S.A, pp. 262-262.

Lockhart, B. E. L. & Olszewski, N. E. (1993). Serological and genomic heterogeneity of *Banana streak badnavirus*: implication for virus detection in *Musa* germplasm. *Breeding Banana and plantain for resistance to disease and pest, J Genry, ed Montpellier France: INIBAP*, 105-113.

Lower, R. (1999). The pathogenic potential of endogenous retroviruses: facts and fantasies. *Trends Microb* **7**, 350-356.

Lower, R., Lower, J. & Kurth, R. (1996). The viruses in all of us: characteristics and biological significance of human endogenous retrovirus sequences. *Proc Natl Acad Sci U S A* **93**, 5177-5184.

Malik, H. S. & Eickbush, T. H. (2001). Phylogenetic analysis of ribonuclease H domains suggests a late, chimeric origin of LTR retrotransposable elements and retroviruses. *Genome Res* **11**, 1187-1197.

Malik, H. S., Henikoff, S. & Eickbush, T. H. (2000). Poised for contagion: evolutionary origins of the infectious abilities of invertebrate retroviruses. *Genome Res* **10**, 1307-1318.

Maori, E., Tanne, E. & Sela, I. (2007). Reciprocal sequence exchange between non-retro viruses and hosts leading to the appearance of new host phenotypes. *Virology* **362**, 342-349.

Matzke, M., Gregor, W., Mette, M. F., Aufstatz, W., Kanno, T., Jakowitsch, J. & Matzke, A. J. M. (2004). Endogenous pararetroviruses of allotetraploid *Nicotiana tabacum* and its diploid progenitors, *N. sylvestris* and *N. tomentosiformis. Biol J Linn Soc* **82**, 627-638.

McKenzie, N., Wen, L. Y. & Dale, P. J. (2002). Tissue-culture enhanced transposition of the maize transposable element Dissociation in *Brassica oleracea* var. 'Italica'. *Theor Appl Genet* **105**, 23-33.

Medberry, S. L., Lockhart, B. E. L. & Olszewski, N. E. (1990). Properties of *Commelina yellow mottle virus*'s complete DNA sequence, genomic discontinuities and transcript suggest that it is a pararetrovirus. *Nucleic acids research* **18**, 5505-5513.

Mette, M. F., Kanno, T., Aufsatz, W., Jakowitsch, J., van der Winden, J., Matzke, M. A. & Matzke, A. J. M. (2002). Endogenous viral sequences and their potential contribution to heritable virus resistance in plants. *Embo J* **21**, 461-469.

Meyer, J. B., Kasdorf, G. G. F., Nel, L. H. & Pietersen, G. (2008). Transmission of activated-episomal Banana streak OL (badna)virus (BSOLV) to cv. Williams banana (*Musa* sp.) by three mealybug species. *Plant Disease* **92**, 1158-1163.

Mitchell, R. S., Beitzel, B. F., Schroder, A. R., Shinn, P., Chen, H., Berry, C. C., Ecker, J. R. & Bushman, F. D. (2004). Retroviral DNA integration: ASLV, HIV, and MLV show distinct target site preferences. *Plos Biol* **2**, E234.

Murad, L., Bielawski, J. P., Matyasek, R., Kovarik, A., Nichols, R. A., Leitch, A. R. & Lichtenstein, C. P. (2004). The origin and evolution of geminivirus-related DNA sequences in *Nicotiana*. *Heredity* **92**, 352-358.

Narezkina, A., Taganov, K. D., Litwin, S., Stoyanova, R., Hayashi, J., Seeger, C., Skalka, A. M. & Katz, R. A. (2004). Genome-wide analyses of *Avian sarcoma virus* integration sites. *Journal of virology* **78**, 11656-11663.

Nash, H. A. (1981). Integration and excision of bacteriophage lambda: the mechanism of conservation site specific recombination. *Annual review of genetics* **15**, 143-167.

Ndowora, T., Dahal, G., LaFleur, D., Harper, G., Hull, R., Olszewski, N. E. & Lockhart, B. (1999). Evidence that *badnavirus* infection in *Musa* can originate from integrated pararetroviral sequences. *Virology* **255**, 214-220.

Nethe, M., Berkhout, B. & van der Kuyl, A. (2005). Retroviral superinfection resistance. *Retrovirology* **2**, 52.

Neumann, K. (2003). New Guinea: A cradle of agriculture. *Science* **301**, 180-181.

Noreen, F., Akbergenov, R., Hohn, T. & Richert-Pöggeler, K. R. (2007). Distinct expression of endogenous *Petunia vein clearing virus* and the DNA transposon dTph1 in two *Petunia hybrida* lines is correlated with differences in histone modification and siRNA production. *Plant J* **50**, 219-229.

Nwakanma, D. C., Pillay, M., Okoli, B. E. & Tenkouano, A. (2003). Sectional relationships in the genus *Musa* L. inferred from the PCR-RFLP of organelle DNA sequences. *Theor Appl Genet* **107**, 850-856.

Paces, J., Pavlicek, A. & Paces, V. (2002). HERVd: database of human endogenous retroviruses. *Nucleic acids research* **30**, 205-206.

Pahalawatta, V., Druffel, K. & Pappu, H. (2008a). A new and distinct species in the genus *Caulimovirus* exists as an endogenous plant pararetroviral sequence in its host, *Dahlia variabilis*. *Virology* **376**, 253-257.

Pahalawatta, V., Druffel, K. L., Wyatt, S. D., Eastwell, K. C. & Pappu, H. R. (2008b). Genome structure and organization of a member of a novel and distinct species of the genus *Caulimovirus* associated with dahlia mosaic. *Arch Virol* **153**, 733-738.

Pal, A., Chakrabarti, A. & Basak, J. (2007). New motifs within the NB-ARC domain of R proteins: Probable mechanisms of integration of geminiviral signatures within the host species of *Fabaceae* family and implications in conferring disease resistance. *J Theor Biol* **246**, 564-573.

Pearson, M. N. & Rohrmann, G. F. (2002). Transfer, incorporation, and substitution of envelope fusion proteins among members of the *Baculoviridae*, *Orthomyxoviridae*, and *Metaviridae* (insect retrovirus) families. *Journal of virology* **76**, 5301-5304.

Puchta, H. (2005). The repair of double-strand breaks in plants: mechanisms and consequences for genome evolution. *J Exp Bot* **56**, 1-14.

Raoult, D., Audic, S., Robert, C., Abergel, C., Renesto, P., Ogata, H., La Scola, B., Suzan, M. & Claverie, J. M. (2004). The 1.2-megabase genome sequence of mimivirus. *Science* **306**, 1344-1350.

Raoult, D. & Forterre, P. (2008). Redefining viruses: lessons from *Mimivirus*. *Nat Rev Micro* **6**, 315-319.

Rattanadechakul, W. & Webb, B. A. (2003). Characterization of *Campoletis sonorensis ichnovirus* unique segment B and excision locus structure. *J Insect Physiol* **49**, 523-532.

Richert-Pöggeler, K. R., Noreen, F., Schwarzacher, T., Harper, G. & Hohn, T. (2003). Induction of infectious petunia vein clearing (pararetro) virus from endogenous provirus in petunia. *Embo J* **22**, 4836-4845.

Richert-Pöggeler, K. R. & Shepherd, R. J. (1997). *Petunia vein-clearing virus*: a plant pararetrovirus with the core sequences for an integrase function. *Virology* **236**, 137-146.

Sabot, F. & Schulman, A. H. (2006). Parasitism and the retrotransposon life cycle in plants: a hitchhiker's guide to the genome. *Heredity* **97**, 381-388.

Safar, J., Noa-Carrazana, J. C., Vrana, J., Bartos, J., Alkhimova, O., Sabau, X., Simkova, H., Lheureux, F., Caruana, M. L., Dolezel, J. & Piffanelli, P. (2004). Creation of a BAC resource to study the structure and evolution of the banana (*Musa balbisiana*) genome. *Genome* **47**, 1182-1191.

Sakharkar, K. R., Dhar, P. K. & Chow, V. T. K. (2004). Genome reduction in prokaryotic obligatory intracellular parasites of humans: a comparative analysis. *Int J Syst Evol Microbiol* **54**, 1937-1941.

Savary, S., Beckage, N., Tan, F., Periquet, G. & Drezen, J. M. (1997). Excision of the polydnavirus chromosomal integrated EP1 sequence of the parasitoid wasp *Cotesia congregata* (*Braconidae*, *Microgastinae*) at potential recombinase binding sites. *J Gen Virol* **78 (Pt 12)**, 3125-3134.

Schmidt, O., Theopold, U. & Strand, M. (2001). Innate immunity and its evasion and suppression by hymenopteran endoparasitoids. *Bioessays* **23**, 344-351.

Schroder, A. R., Shinn, P., Chen, H., Berry, C., Ecker, J. R. & Bushman, F. (2002). HIV-1 integration in the human genome favors active genes and local hotspots. *Cell* **110**, 521-529.

Simmonds, N. W., (ed) (1962). *The evolution of the bananas, Longmans Green (Ed)*. London.

Simmonds, N. W. & Shepherd, K. (1955). The taxonomy and origins of the cultivated bananas. *J Linn Soc Bot* **LV**, 302-312.

Slotkin, R. K. & Martienssen, R. (2007). Transposable elements and the epigenetic regulation of the genome. *Nat Rev Genet* **8**, 272-285.

Song, S. U., Gerasimova, T., Kurkulos, M., Boeke, J. D. & Corces, V. G. (1994). An env-like protein encoded by a *Drosophila* retroelement: evidence that gypsy is an infectious retrovirus. *Genes Dev* **8**, 2046-2057.

Staginnus, C., Gregor, W., Mette, M. F., Teo, C. H., Borroto-Fernandez, E. G., Machado, M. L., Matzke, M. & Schwarzacher, T. (2007). Endogenous pararetroviral sequences in tomato (*Solanum lycopersicum*) and related species. *BMC Plant Biol* **7**, 24.

Staginnus, C. & Richert-Pöggeler, K. R. (2006). Endogenous pararetroviruses: two-faced travelers in the plant genome. *Trends Plant Sci* **11**, 485-491.

Suttle, C. A. (2005). Viruses in the sea. *Nature* **437**, 356-361.

Suzan-Monti, M., La Scola, B. & Raoult, D. (2006). Genomic and evolutionary aspects of *Mimivirus*. *Virus Res* **117**, 145-155.

Tanne, E. & Sela, I. (2005). Occurrence of a DNA sequence of a non-retro RNA virus in a host plant genome and its expression: evidence for recombination between viral and host RNAs. *Virology* **332**, 614-622.

Temin, H. M. (1980). Origin of retroviruses from cellular moveable genetic elements. *Cell* **21**, 599-600.

Terzian, C., Pelisson, A. & Bucheton, A. (2001). Evolution and phylogeny of insect endogenous retroviruses. *BMC evolutionary biology* **1**, 3.

Tinsley, C. R., Bille, E. & Nassif, X. (2006). Bacteriophages and pathogenicity: more than just providing a toxin? *Microbes Infect* **8**, 1365-1371.

Turnbull, M. & Webb, B. (2002). Perspectives on polydnavirus origins and evolution. In *Adv Virus Res*, pp. 203-254: Academic Press.

Ude, G., Pillay, M., Nwakanma, D. & Tenkouano, A. (2002). Analysis of genetic diversity and sectional relationships in *Musa* using AFLP markers. *Theor Appl Genet* **104**, 1239-1245.

Uma, S., Siva, S. A., Saraswathi, M. S., Durai, P., Sharma, T., Singh, D. B., Selvarajan, R. & Sathiamoorthy, S. (2005). Studies on the origin and diversification of Indian wild banana (*Musa balbisiana*) using arbitrarily amplified DNA markers. *J Hortic Sci Biotech* **80**, 575-580.

Vaillant, I., Schubert, I., Tourmente, S. & Mathieu, O. (2006). MOM1 mediates DNA-methylation-independent silencing of repetitive sequences in *Arabidopsis*. *Embo Reports* **7**, 1273-1278.

Wagner, P. L. & Waldor, M. K. (2002). Bacteriophage control of bacterial virulence. *Infection and immunity* **70**, 3985-3993.

Wardlaw, C. W. (1972). *Banana diseases including plantain and abaca*: Longman Group Ltd: London UK.

Wassenegger, M. (2005). The role of the RNAi machinery in heterochromatin formation. *Cell* **122**, 13-16.

Weinbauer, M. G. & Rassoulzadegan, F. (2004). Are viruses driving microbial diversification and diversity? *Environ Microbiol* **6**, 1-11.

Weiss, R. A. (2006). The discovery of endogenous retroviruses. *Retrovirology* **3**.

Wickett, N. J., Zhang, Y., Hansen, S. K., Roper, J. M., Kuehl, J. V., Plock, S. A., Wolf, P. G., dePamphilis, C. W., Boore, J. L. & Goffinet, B. (2008). Functional gene losses occur with minimal size reduction in the plastid genome of the parasitic liverwort *Aneura mirabilis*. *Mol Biol Evol* **25**, 393-401.

Wong, C., Kiew, R., Argent, G., Set, O., Lee, S. K. & Gan, Y. Y. (2002). Assessment of the validity of the sections in *Musa* (Musaceae) using AFLP. *Ann Bot-London* **90**, 231-238.

Wu, X., Li, Y., Crise, B. & Burgess, S. M. (2003). Transcription start regions in the human genome are favored targets for MLV integration. *Science* **300**, 1749-1751.

Wyder, S., Tschannen, A., Hochuli, A., Gruber, A., Saladin, V., Zumbach, S. & Lanzrein, B. (2002). Characterization of *Chelonus inanitus polydnavirus* segments: sequences and analysis, excision site and demonstration of clustering. *J Gen Virol* **83**, 247-256.

Yang, I. C., Hafner, G. J., Revill, P. A., Dale, J. L. & Harding, R. M. (2003). Sequence diversity of South Pacific isolates of *Taro bacilliform virus* and the development of a PCR-based diagnostic test. *Arch Virol* **148**, 1957-1968.

Zeidan, M., Sikron, N., Cohen, J. & Gera, A. (2000). Improved detection of *Petunia vein clearing caulimovirus*. *Hortscience* **35**, 1279-1282.

Zhang, J. Z. (2003). Evolution by gene duplication: an update. *Trends Ecol Evol* **18**, 292-298.

Zohary, D. & Hopf, M. (2000). *Domestication of plants in the old world, 3rd edn*: Oxford University Press, London UK.

ANNEXES

Annexe 1 - Article 5 : "Exploring the banana streak viruses - *Musa* sp. pathosystem: how does it work?"

P. Gayral, F. Lheureux, J.C. Noa-Carrazana, M. Lescot, P. Piffanelli, F. Carreel, C. Jenny and M.L. Iskra-Caruana. Exploring the banana streak viruses - *Musa* sp. pathosystem: how does it work? (2007), Proceedings of ISHS/ProMusa symposium White River, South Africa September 10-14. Recent advances in banana crop protection for sustainable production and improved livelihoods. **Acta Horticultura** - à paraitre.

P. Gayral[1], F. Lheureux[1], J.C. Noa-Carrazana[2a], M. Lescot[2b], P. Piffanelli[2c], F. Carreel[3], C. Jenny[3] and M.L. Iskra-Caruana[1]

[1]Centre de Coopération Internationale en Recherche Agronomique pour le Développement (CIRAD), BIOS UMR 385 BGPI TA A554/K, Campus International de Baillarguet, F-34398 Montpellier Cedex 5, France
[2]CIRAD, BIOS UMR DAP TA40/03, Av. Agropolis, F-34398, Montpellier Cedex 5, France
[2a]Laboratorio Instituto de Biotecnología y Ecología Aplicada, Univ. Veracruzana, Av. Culturas Veracruzanas 101, Col E. Zapata, CP 91090, Xalapa, Ver., México
[2b]Present address: Structural and Genomic Information Laboratory (IGS), C.N.R.S. UPR 2589, Institute of Structural Biology and Microbiology (IBSM), Parc Scientifique de Luminy, 163 avenue de Luminy, F-13288, Marseille Cedex 9, France
[2c]Present address: Rice Genomics group, AgBiotech Research Centre, Parco Tecnologico Padano, Via Einstein, Località Cascina Codazza, 26900 Lodi, Italy
[3]CIRAD, BIOS UPR 75, Station de Neufchâteau, 97130 Capesterre Belle-Eau, Guadeloupe (FWI)

Abstract

Banana streak viruses (BSVs) are double-stranded DNA pararetroviruses causing banana streak disease. Recently, numerous outbreaks of the disease occurred in many banana-producing areas in interspecific hybrids (*Musa acuminata* x *Musa balbisiana*) originating from virus-free parents. These infections correlated with the presence of endogenous banana streak viruses, viral DNA sequences integrated in the *M. balbisiana* genome only. Although integration is not needed for the viral replication cycle, some viral integrants may become infected under stress conditions by reconstituting a replication-competent genome after homologous recombination events. Even though the wild *M. balbisiana* 'Pisang Klutuk Wulung' (PKW) harbours infectious endogenous BSVs (eBSVs), it is resistant to BSVs. We characterized the genetic and genomic endogenous viral organization of three BSV species in PKW in order to determine the species responsible for the viral expression in the interspecific F1 progeny.

Keywords: Endogenous pararetroviruses, *Musa* genome, BEL factor

INTRODUCTION

Banana streak viruses (BSVs) are plant pararetroviruses, belonging to the genus *Badnavirus* in the family of the *Caulimoviridae* (Fauquet et al., 2005). They are bacilliform viruses containing double-stranded DNA of 7.4 kbp with 3 ORFs. They display important serological and molecular variability with more than 20% nucleotide differences in the most conserved RT-Rnase H region (E. Muller, pers. commun.). For this reason, BSVs are considered as different badnavirus species. All BSV species are capable of causing banana streak disease, an economically important viral disease of banana. The characteristic symptoms of the disease are yellow streaks on leaves that turn necrotic, splitting of the pseudostem and, in severe cases, cigar leaf necrosis leading to the death of the banana plant.

As with several other plants (Hull et al., 2000; Staginus et Richert- Pöggeler, 2006), the genome of banana contains endogenous banana streak viruses, even though integration is not an essential step in the replication cycle of these viruses. Two types of BSV integrants exist in banana. Non-functional sequences are present in both *Musa acuminata* (denoted A) and *Musa balbisiana* (denoted B). Integrants of the second type are restricted to the *M. balbisiana* genome and contain the complete viral genome. It is assumed that these may become infectious by reconstituting a complete replication-competent viral genome (Ndowora et al., 1999; Geering et al., 2000; Lheureux et al., 2003).

BSVs are horizontally transmitted by mealybugs and infected suckers, but can also result from the activation of eBSVs present in the banana genome (Fargette et al., 2006). The increasing numbers of records of BSV outbreaks observed in banana breeding lines and micropropagated interspecific banana hybrids worldwide result from such eBSVs present in the *M. balbisiana* genome producing infectious virions under stress conditions (Lockhart and Jones, 2000).

So far, it has not been possible to identify which banana plants having integrated BSV sequences will release virions when submitted to a stress. Therefore, it is difficult to predict and prevent outbreaks, and to manage the high inoculum levels that occur when several plants become infected at the same time. Epidemics of BSVs remain, therefore, very difficult to control. In addition, BSVs have become the main viral constraint for the safe movement of banana germplasm, and for genetic improvement efforts. Scientists affiliated with the *Centre de Coopération Internationale en Recherche Agronomique pour le Développement* (CIRAD) have investigated the mechanisms underlaying activation.

DISEASE EXPRESSION

Three widespread BSVs, *Banana streak Obino l'Ewai virus* (BSOLV), *Banana streak Imové virus* (BSImV) and *Banana streak Goldfinger virus* (BSGFV), are known to occur as infectious integrants in the *M. balbisiana* genome. Although infectious, their presence alone is not enough to induce infection. The context that may trigger the episomal expression of EPRVs include the process of genetic hybridization,

micropropagation by in-vitro culture (Dallot et al., 2001) and abiotic stresses, such as temperature differences and water stress.

Disease expression was characterized by conducting interspecific genetic crosses. The diploid virus-free *M. acuminata* 'IDN 110' was doubled by colchicine and crossed with the female diploid virus-free *M. balbisiana* 'Pisang Klutuk Wulung' (PKW) to produce an interspecific triploid (BAA) F1 population. We observed that half of the progenies was virus free while the other half was infected. External contamination was impossible. We assumed a Mendelian segregation (50:50) of the disease after checking viral particles by immunosorbent electron microscope (ISEM) and Multiplex IC-RT-PCR (Leprovost et al., 2005).

In order to characterize the genetic factors involved in disease expression, AFLP markers co-segregating with the presence of virions based on bulk segregant analysis were sought. Ten AFLP markers were identified, all of them present only in the *M. balbisiana* parent. Seven markers segregated with 'the presence of virions' while three, located on homologous locus, segregated with the 'absence of the virions' (Lheureux et al, 2003).

The viral distribution of BSOLV, BSImV and BSGfV among the BAA F1 progeny as a monogenic allelic system conferring the role of carrier to the diploid *M. balbisiana* parent was characterized. Segregation analysis of AFLP markers using the Map Maker® software allowed the construction of a genetic map of the locus, including the BSV expressed locus (BEL), the genetic factor involved in disease expression. BSOLV and BSImV appeared in most infected hybrids depending on BEL regulation, while BSGFV were restricted to only half of the infected hybrids and subordinated by BEL (Iskra-Caruana et al., 2003). Thus, disease expression seems to have a genetic origin.

GENETIC STRUCTURE OF INFECTIOUS ENDOGENOUS BSVs (eBSVs)

The genomic structure of endogenous BSVs in banana was investigated. Three BAC libraries from 'Petite Naine' (AAA, Cavendish subgroup), 'Calcutta 4' (AAw, *M. acuminata* ssp. *burmannicoides*) and PKW were made and explored for the pattern of integration of infectious endogenous BSVs (eBSVs). A set of different viral probes was tested, representing each time the complete BSOLV, BSImV, BSGFV and *Banana streak Mysore Virus* (BSMyV) genome. The presence of BEL was tested using AFLP markers as probes (Safar et al., 2004). BSV-positive BAC clones were characterised by RFLP fingerprint approaches. The analysis showed that the four BSV species represent low-copy loci and that their integrations are specific to the PKW *M. balbisiana* genome. PKW contains at least seven different BSVs integrants: three for BSOLV, one for BSImV, two for BSGFV and one for BSMyV. One BAC clone for each group of integrants has been sequenced, and four BAC clones have been fully annotated after sequencing: one for BSOLV, two for BSGFV and recently one for BSImV. Each integrant is composed of complex back-to-back viral sequences representing more than a whole BSV genome.

Even if PKW is known to be virus free, it harbours eBSVs since virions of at least three BSV species are observed in its progeny after interspecific genetic crosses. Attempts were made to identify the eBSVs responsible for the viral expression of each integrated BSV species. Three kinds of BSOLV integrated sequences are present in PKW. Only type 1 is annotated, showing successions of partial viral sequences back to back representing at least one total viral genome. The annotation for the two other types is in progress and seems more complicated. But three different PCR markers designed in vital zones of the BSOLV genome (CP and RT-RNAse H genes and IGR intergenic region, ORFs 1 and 2) allow the type 1 BAC clone to be distinguished from the others. In the progeny, all infected banana plants display the same PCR pattern, whilst healthy plants never react. This pattern is also observed in PKW. Therefore, it was concluded that the integrant of type 1 is involved in the release of virions.

Studies on the BSGfV integrated sequences are more complete thanks to access to the overall sequences of both kinds of integrants, named 7 and 9. They appear very similar and show 99.7 % identity. The unique difference is in the presence of an insertion of 3 kbp in type 9. A conserved synteny of the *Musa* genes around the endogenous BSVs is also observed suggesting that endogenous BSV 7/9 could be an allelic insertion. We developed genotyping molecular markers (PCR, PCR-RFLP) to distinguish the two types and analysed their segregation in the AAB F1 progeny. BSGfV integrants were found to be allelic, located at the same locus. It is concluded that the integrated sequences of BSGfV in PKW result from one integration event (Gayral et al., 2008.). In order to determine which type 7 or type 9 of endogenous BSV is infectious, primers specific to the circular viral form were developed. All infected hybrids harbour type 7 of BSGfV integrants.

CONCLUSION

Little is currently known about the actual mechanisms underlying the genetic expression of eBSVs and their regulation. The role of methylation, described as a possible regulator for triggering disease expression in other pathosystems, was investigated. Differential cytosine methylation patterns were searched for in healthy and diseased F1 BAA hybrids using the SD-AFLP/MSAP technique. The role of chromosomal rearrangements was also investigated through a PCR-based analysis of both genomic DNA extracted from the progeny and BSV positive BAC clones of the *M. balbisiana* parent PKW. Among the thirteen DNA fragments obtained by SD-AFLP/MSAP, one corresponds to an AFLP marker closely located to the BEL locus. Differential PCR patterns were observed, depending on the strain-specific primers used, covering distinct parts of the BSV genome and suggesting chromosomal rearrangements in diseased hybrids. This is corroborated by the analysis of BAC clones of the PKW parent. Nevertheless, there is no convincing evidence that methylation plays a major role in the activation of eBSVs.

The complete annotation of the other BAC clones is currently in progress, and a similar study will be performed in order to identify which type of insertion is responsible for BSV expression in the progeny.

ACKNOWLEDGEMENTS

We thank Ms. Kozue Kamiya, Dr. Hiroyuki Kanamori, Dr. Takashi Matsumoto, and Dr. Takuji Sasaki for performing the sequencing of the two BAC clones (71C19 and 94I16) at the National Institute of Agrobiological Sciences (NIAS) in Japan. We acknowledge the financial support to J.C. Noa-Carrazana from the Project Agropolis II coordinated by Agropolis, Montpellier and Bioversity-France in France, and CINVESTAV in Mexico for sequencing of one BAC clone, and the access to material facilitated by the Global *Musa* Genomics Consortium. Philippe Gayral is supported by a 'CIRAD/Région Languedoc-Roussillon' PhD grant.

Literature Cited

Dallot, S., Acuña, P., Rivera, P., Ramirez, P., Cote, F., Lockhart, B.E.L. and Caruana, M.L. 2001. Evidence that the proliferation stage of micropropagation procedure is determinant in the expression of *Banana streak virus* integrated into the genome of the FHIA 21 hybrid (*Musa* AAAB). Arch. Virol. 146(11):2179-2190.

Fargette, D., Konate, G., Fauquet, C., Muller, E., Peterschmitt, M. and Thresh, J.M. 2006. Molecular ecology and emergence of tropical plant viruses. Annu. Rev. of Phytopathology 44:235-260.Fauquet, C.M., Mayo, M.A., Maniloff, J., Desselberger, U. and Ball, L.A. 2005. Virus taxonomy: 8th report of the International Committee of the Taxonomy of Viruses. Elsevier Academic Press, Amsterdam.

P. Gayral, J.-C. Noa-Carrazana, M. Lescot, F. Lheureux, B. E. L. Lockhart, T. Matsumoto, P. Piffanelli and M.-L. Iskra-Caruana. (2008) A single *Banana streak virus* integration event in the banana genome as the origin of infectious endogenous pararetrovirus (EPRV) Journal of Virology vol 82 N°13 pp 6697-6710

Geering, A.D.W., McMichael, L.A., Dietzgen, R.G., and Thomas, J.E., 2000. Genetic diversity among *Banana streak virus* isolates from Australia. Phytopathology 90:921-927.

Hull, R., Harper, G., Lockhart, B.E.L. 2000. Viral sequences integrated into plant genome. Trends in Plant Science 5(9):362-365.

Iskra Caruana M.L., Lheureux F., Noa-Carrazana J.C., Piffanelli P., Carreel F., Jenny C., Laboureau N., Lockhart B.E.L. 2003. Unstable balance of relation between pararetrovirus and its host plant: the BSV-EPRV banana pathosystem. Abstracts : EMBO Workshop Genomic Approaches in Plant Virology, May 28-31. Keszthely, Hungary. p 8.

Le Provost, G., Iskra-Caruana, M.L., Acina, I. and Teycheney, P.Y. (2006). Improved detection of episomal Banana streak viruses by multiplex immunocapture PCR. J. Virol. Methods 137(1):7-13

Lheureux, F., Carreel, F., Jenny, C., Lockhart, B.E.L. and Iskra-Caruana, M.L. 2003. Identification of genetic markers linked to Banana streak disease expression in inter-specific *Musa* hybrids. Theor. Appl. Genetic 106:594-598.

Lockhart, B.E.L and Jones, D.R. 2000. Banana streak virus. In: D.R. Jones (ed.), Diseases of banana, abaca and enset. CABI, Wallingford. p. 263-274.

Ndowora, T., Dahal, G., LaFleur, D., Harper, G., Hull, R., Olzsewski, N. and Lockhart, B.E.L. 1999. Evidence that badnavirus infection in *Musa* can originate from integrated pararetroviral sequences. Virology 255:214-220.

Staginnus, C. and Richert- Pöggeler, K.R. 2006. Endogenous pararetroviruses: two-faced travelers in the plant genome. Trends in Plant Science 11(10):485-491.

Safar, J., Noa-Carrazana, J.C., Vrana, J., Bartos, J., Alkhimova, O., Sabau, X., Simkova, H., Lheureux, F., Caruana, M.L., Dolezel, J. and Piffanelli, P. 2004. Creation of a BAC resource to study the structure and evolution of the banana (*Musa balbisiana*) genome. Genome 47(6):1182-1191.

Annexe 2 - Article 6 : "How to Control and Prevent the Spread of Banana Streak Disease when the Origin could be Viral Sequences Integrated in the Banana Genome"

M.L. Iskra-Caruana, **P. Gayral**, S. Galzi, N. Laboureau. How to control and prevent the spread of *banana streak virus* (BSV) when the origin could be viral sequences integrated in *Musa* genome? (2007), Proceedings of ISHS/ProMusa symposium White River, South Africa September 10-14. Recent advances in banana crop protection for sustainable production and improved livelihoods. **Acta Horticultura** - à paraitre.

M.L. Iskra-Caruana, Philippe Gayral, S. Galzi and N. Laboureau

Centre de Coopération Internationale en Recherche Agronomique pour le Développement (CIRAD), BIOS UMR 385 BGPI TA A554/K, Campus International de Baillarguet, F-34398 Montpellier Cedex 5, France

Abstract

Banana streak viruses are among the most widely distributed viruses of banana and are responsible for banana streak disease. Natural field spread occurs by either mealybugs or use of infected planting material, such as suckers. Banana streak viruses are pararetroviruses belonging to the genus *Badnavirus*, in the family Caulimoviridae. Like all members of the *Badnavirus* genus, they have bacilliform virions, 30 × 150 nm in size, and a circular dsDNA genome of 7.4 kbp. Fifteen years ago, an increasing number of outbreaks of banana streak disease were reported worldwide. Many occurred in banana breeding lines and micropropagated inter-specific banana hybrids. The origin of infections in new hybrids and tissue-cultured plants was linked to the presence of viral DNA sequences integrated into the *Musa balbisiana* genome. Although integration is not an essential step in the replication cycle of pararetroviruses, it is assumed that under stress conditions some endogenous banana streak viruses could become infectious by reconstituting a complete replication-competent viral genome. Several serological and molecular tools have been developed to detect either virions or endogenous banana streak viruses. Their specificity and potential to prevent and control outbreaks of banana streak disease is discussed.

Keywords: Banana streak virus, diagnostic, endogenous pararetroviruses, *Musa*

INTRODUCTION

Banana streak disease is caused by a complex of banana streak viruses. All are mealybug-transmitted plant bacilliform pararetroviruses belonging to the genus *Badnavirus* within the family Caulimoviridae (Fauquet et al., 2005; Lockhart, 1986). Their genome consists of a double-stranded, non-covalently linked circular DNA of 7.4-kb and they have a wide serological and molecular variability. Banana streak virus infections cause characteristic chlorotic and necrotic streaks on leaves with highly susceptible banana cultivars developing more severe symptoms, such as pseudostem splitting, which eventually leads to the death of infected plants (Lockhart and Jones, 2000; Fargette et al., 2006). Banana streak viruses are among the most widely distributed viruses of banana and have never been considered a serious threat (Daniells et al., 2001) until recently (Fargette et al., 2006) when a number of outbreaks of the disease occurred in promising virus-free banana breeding lines and micropropagated inter-specific *Musa* hybrids (Lheureux et al., 2003; Dallot et al., 2001). Such infections were correlated with the presence of banana streak virus sequences integrated into the nuclear genome of *M. balbisiana*. Integrated virus sequences exist in *Musa acuminata* and *Musa balbisiana* although the majority appear to be inactive due to premature stop codons, frame-shift mutations and perhaps incomplete genomes (Geering et al., 2005a). From field experience, only endogenous banana streak viruses (eBSVs) from *M. balbisiana* appear capable of being reconstituted into a complete replication-competent viral genome (Harper et al., 1999a; Ndowora et al., 1999; Dallot et al., 2001; Lheureux et al., 2003). It is thought that epidemics of banana streak arise as a consequence of both the activation of eBSVs and mealybug spread of exogenous forms of the virus.

Diagnostics for banana streak viruses are difficult because of the broad sequence diversity in the banana streak virus complex and the existence of sequences integrated in the banana genome. Their presence in the banana genome hampers the detection of cognate episomal viruses by PCR, since PCR amplifies viral DNA of both viral particles and integrated viral sequences, which leads to false positives (Harper et al., 1999b; Yang et al., 2003). Banana streak viruses are today an important constraint to banana germplasm movement, genetic improvement and mass propagation. Several serological and molecular techniques have been developed so far and detect either virions or eBSVs or both. A relevant generic diagnostic for all banana streak viruses is needed. We have compared several immunocapture PCR assays to determine which is best with regards to sensitivity and ability to detect a broad range of banana streak viruses.

MATERIAL AND METHODS

Banana plants were grown under insect-proof conditions in a tropical glasshouse at CIRAD in Montpellier, France. Individual banana streak infections of cv. Cavendish plants were obtained by mealybug transmission using *Planococcus citri*. These infections corresponded to each following banana streak virus species used in this study and kindly provided by B.E.L. Lockhart.: *Banana streak Obino L'ewai virus* (BSOLV), *Banana streak Goldfinger virus* (BSGFV), *Banana streak Imové virus* (BSImV),

Banana streak Cavendish virus (BSCavV), *Banana streak acuminata Vietnam virus* (BSAcVNV) and *Banana streak Mysore virus* (BSMyV).

Partial purification of the viruses was realised by grinding 10g of banana leaf tissue ground in liquid nitrogen. The powder was extracted in 20ml of 200mM Tris-HCl pH 7.4 containing 1% of Na_2SO_3 and then was filtered through cheesecloth. After centrifugation at 13,000-14,000 x g for 10-15 min, 1ml of triton X100 at 33% was added to the filtrate before layering it over 6 ml of 30% of sucrose in 15mM Tris-HCl pH 7.4. The centrifugation was done during 60min in a Beckman 50.2 rotor at 30,000rpm (109,000 x g max). After centrifugation and elimination of the supernatant, the sides of the tube were rinsed to eliminate residual triton X100. The pellet was resuspended in 200µl of 1.5mM Tris HCl pH 7.4 and clarified by centrifugation at 8000g for 5min before its transfert into microtubes to be kept at 4°C.

All immunocapture PCRs were performed using the protocol of Le Provost et al. (2006). All PCR mixtures contained 50 ng of DNA (prepared using a Qiagen Plant Genomic DNA kit), 1 × Taq DNA polymerase buffer of20 mM Tris–HCl (pH 8.4), 50 mM KCl), 0.1 mM each dNTP, 1.5 mM $MgCl_2$, 10 pmol each of reverse and forward primers and 1U *Taq* DNA polymerase (Eurogentec, Seraing, Belgium) in a final volume of 25 µL. DNA was amplified following specific amplification cycles related to the sets of primers used (Table 1 and 2). Amplicons were separated on a 1.5-% agarose gel in 0.5 × TBE (45 mM Tris-borate, 1 mM EDTA, pH 8), stained with ethidium bromide, and the expected bands visualized on a UV transilluminator.

Table 1. Nucleotide sequence of primers used in IC-PCR and PCR experiments to test for banana streak viruses

Name of primer	Primer sequence (5'-3')	Size of PCR product (bp)	Target
RDOL-F1[1]	ATCTGAAGGTGTGTTGATCAATGC	522	BSV-Ol
RD-R1	GCTCACTCCGCATCTTATCAGTC		
GF-F1[1]	ACGAACTATCACGACTTGTTCAAGC	476	BSV-Gf
GF-R1	TCGGTGGAATAGTCCTGAGTCTTC		
Badna 1A[2]	TAAAAGCACAGCTCAGAACAAACC	589	Badnavirus
Badna 4'	CTCCGTGATTTCTTCGTGGTC		
Badna RP[3]	CCA YTT RCA IAC ISC ICC CCA ICC	570	Badnavirus
Badna FP	ATGCCITTYGGIAARAAYGCICC		
AGMI 025[4]	TTAAAGGTGGGTTAGCATTAGG	248*	STMS
AGMI 026	TTTGATGTCACAATGGTGTTCC		

*: size of PCR fragment amplified from *M. balbisiana* genomic DNA.
[1]Geering et al., 2000; [2]Geering et al., 2005; [3]Yang et al., 2003; [4]Lagoda et al., 1998.

All multiplex IC-PCR and multiplex PCR assays were undertaken using the common mix described above. The ratio between the banana streak virus- or badnavirus-specific sets of primers and the STMS primers was 10 pmol: 30 pmol. The amplification cycles were as described in Tables 1 and 2.

Table 2. PCR cycles of amplification for each set of primers used.

RD-F1/RD-R1	GF-F1/GF-R1	Badna 1/4	Badna FP/RP
94°C -3 min	94°C -3 min	min 94°C -3 min	94°C -2 min
30 cycles (94°C- 30s, 58°C-30s, 72°C-30s)	30 cycles (94°C- 30s, 58°C-30s, 72°C-30s)	5 cycles (94°C- 30s, 37°C-30s, 72°C-1 min 30)	35 cycles (94°C- 30s, 55°C-30s, 72°C-30s)
72°C – 5 min	72°C – 5 min	30 cycles (94°C- 30s, 50°C-30s, 72°C-30s)	72°C – 10 min
		72°C – 10 min	

RESULTS AND DISCUSSION

Capacities of Existing Tools for Detecting BSV

Several sets of primers are available to detect banana streak virus sequences (Table 1), some of which are specific to each banana streak virus while others are badnavirus-generic. The sensitivities of these PCR assays to detect banana streak virus virions were compared by IC-PCR using the same viral concentration of leaf extracts from banana plants infected by BSOLV and BSGfV. Figure 1 shows that the BSOLV- and BSGfV-specific primers were able to detect virus over a dilution of 1:100,000. The Badna 1A/4' badnavirus-generic primers did not detect BSGFV at any dilution while BSOLV could only be detected down to a dilution of 1:100. Badnavirus RP/FP, the alternative badnavirus-generic primers, were as sensitive as the primers specific to a banana streak virus species, but unfortunately, this set of primers also reacted with the uninfected banana leaf samples, indicating that the immunocapture step did not exclude all unencapsidated DNA.

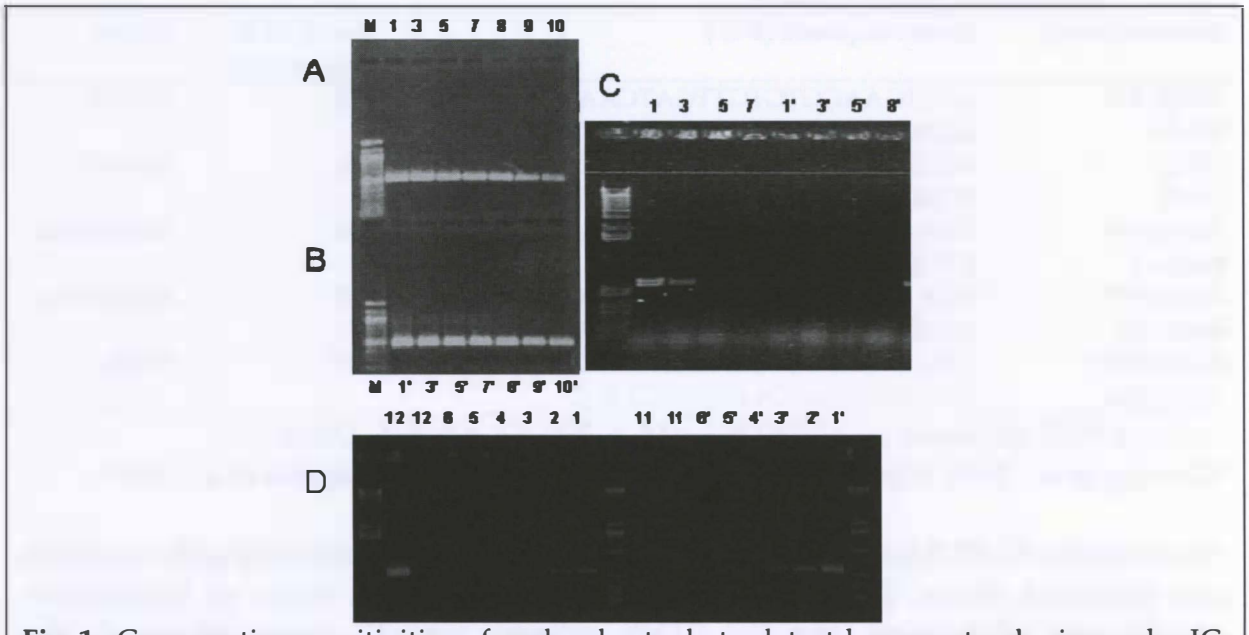

Fig. 1. Comparative sensitivities of molecular tools to detect banana streak viruses by IC-PCR: A and B: BSOLV- and BSGFV-specific primers, respectively; C: BSV-specific primers (Badna 1A/4'); D: Badnavirus-specific primers (Badna FP/FR). Leaf extracts from BSOLV (1 to 10) and BSGFV (1' to 10') infected plants diluted at 1:10, 1:20, 1:100, 1:500, 1:1,000, 1:2,000, 1:5,000, 1:10,000, 1:50,000, 1:100,000; (11) Leaf extracts from virus-free banana plant; (12) Total DNA from the virus-free 'Pisang Klutuk Wulung' (BB). M: 1Kb DNA marker (Invitrogen)

Generic Diagnostic of BSV by IC-PCR

The ability of the badnavirus-generic primers to detect a range of banana streak virus species were assessed against all six different viruses used in this study. The Badna 1A/4' primers well detected all viruses by IC-PCR after partial purification of the viruses (data not shown). All six banana streak viruses were also detected with the badna FP/RP primers (Figure 2), although relative band intensities were different according to banana streak virus species detected. This did not appear to be due to differences in viral concentration since the plant samples are controlled using species-specific sets of primers (data not shown). Strong amplification was observed for BSCavV and only slight reactions with both BSOLV, BSMyV and BSAcVNV (data not shown). Unexpected PCR bands were observed with both virus-free banana samples used as negative controls, cv. Cavendish plants and the wild *M. balbisiana* diploid 'Pisang Klutuk Wulung' (PKW). These reactions result from the binding of residual *Musa* genomic DNA to the walls of tubes or microplates used for IC-PCR leading to the amplification of integrated viral sequences.

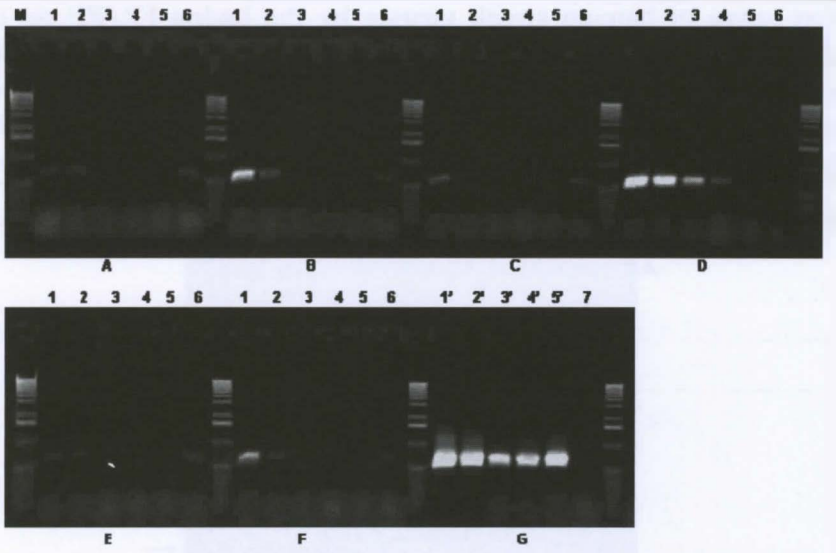

Fig. 2. Detection range of banana streak viruses by the badna FP/RP set of primers by IC-PCR: Banana streak virus-infected banana leaf samples with A: BSOLV; B: BSGFV; C: BSImV; D: BSCavV; E: BSMyV; F: Uninfected banana leaf samples from 'Pisang Klutuk Wulung' (BB); G: Total DNA extract from 'Pisang Klutuk Wulung'.
(1 to 5) Dilutions of banana leaf extract at 1:10, 1:20, 1:100, 1:500, 1:1000; (1' to 5') Dilutions of total DNA 1µl, 0.5µl, 0.2µl, 0.1µl, 0.05µl; (6) Uninfected banana leaf samples from a Cavendish cultivar (AAA); (7) H_2O.
M: 1Kb DNA marker (Invitrogen)

Generic Diagnostic by Multiplex IC-PCR

To alert the user of residual *Musa* genomic DNA while still retaining sensitivity of detection, a multiplex IC-PCR using *Musa* sequence target microsatellite site primers (STMS), as developed by Le Provost et al. (2006), was undertaken. A multiplex PCR was firstly performed in order to test the reactivity of the combination of both STMS and badna FP/RP within the same tube. Several combinations to multiplex the two sets of primers were set up. The best one is illustrated in Figure 3. The two controls were correct: the microsatellite PCR amplification was recorded for both DNA from Cavendish plants and PKW while the

eBSVs amplification was recorded for DNA from PKW only. Surprisingly, BSOLV, BSImV and BSAcVNV, which were detected in previous tests, did not react in this test, whereas the microsatellite amplification is correct at least for two of them.

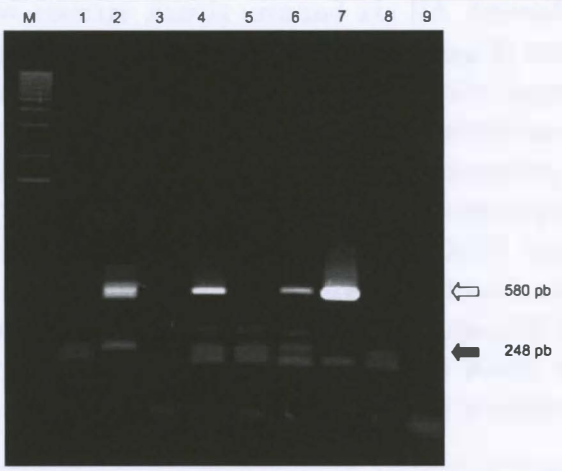

Fig. 3. Detection range of banana streak viruses by the badna FP/RP set of primers and multiplex PCR. (1) Uninfected banana leaf samples from a Cavendish cultivar (AAA); (2) Uninfected banana leaf samples from 'Pisang Klutuk Wulung' (BB); (3 to 8) Banana leaf samples from a Cavendish cultivar (AAA) infected by BSOLV, BSGFV, BSImV, BSMyV, BSAcVNV and BSCavV; (9) H_2O. M: 1Kb DNA marker (Invitrogen)

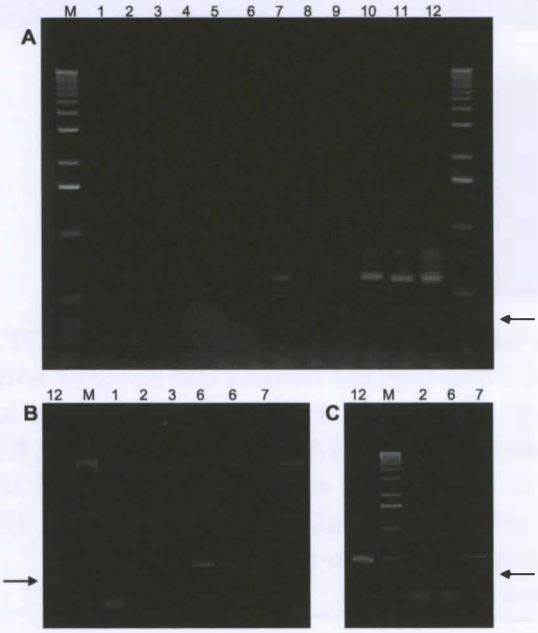

Fig. 4. Multiplex IC-PCR to detect banana streak viruses: A: multiplex of Badna FP/RP and STMS primers; B: multiplex of BSOLV-specific and STMS primers; C: multiplex of BSGFV-specific and STMS primers. Arrows indicate the STMS amplification. Banana leaf samples from (2) uninfected banana plants of a Cavendish cultivar (AAA), (3) 'Pisang Klutuk Wulung' (BB), (4) 'Penkelon' (AAB), (5) 'CRBP 39' (AAAB); (6 to 11) 'Cavendish' infected by BSOLV, BSGFV, BSImV, BSCavV, BSAcVNV; (12) DNA extract from 'Pisang Klutuk Wulung'; 1: H_2O. M: 1Kb DNA marker (Invitrogen)

A multiplex IC-PCR was undertaken with the same multiplex conditions and results confirmed (Figure 4A). Figures 4B and 4C show that the infected leaf samples

used to realize the IC step are infected since a strong amplification was recorded in multiplex IC-PCR using specific sets of primers for both BSOLV and BSGFV, respectively.

DISCUSSION AND CONCLUSIONS

Among the several molecular tools existing to detect banana streak viruses, the species-specific PCR primers are clearly the best for detecting individual species by IC-PCR. However, broader specificity is required and a relevant generic molecular banana streak virus diagnostic does not exist today. Even if the range of detection for the Badna 1A/4' tool is correct, they do not have sufficient sensitivity in IC-PCR unless a partially purified virus sample is used. Partial purification to concentrate the virus in the sample is very time-consuming and requires expensive equipment, which must be available if this process is to be used in the routine diagnosis of banana streak virus. The Badna FP/RP primer set provides greater sensitivity, although this extra sensitivity revealed problems with the immunocapture step, as residual *Musa* genomic DNA could be detected. This last experience reveals the need to systematically develop a multiplex IC-PCR approach in order to allow the detection of false positive resulting from the presence of residual *Musa* genomic DNA containing endogenous banana streak viruses. Multiplex PCR assays containing controls for DNA have to be developed (Le Provost et al., 2006), although such assays are problematic when the badnavirus primers are too degenerate. An alternative is to develop a molecular test specific to the circular viral genome, such as those based on rolling-circle and long PCR amplification. Even if this test works, it is expensive for diagnostic use. The last solution is to design a set of primers which does not react with any of the endogenous banana streak viruses and which only amplifies the circular viral form. This requires a wide knowledge of all patterns of integration existing in banana.

Literature Cited

Dallot, S., Acuña, P., Rivera, P., Ramirez, P., Cote, F., Lockhart, B.E.L. and Caruana, M.L. 2001. Evidence that the proliferation stage of micropropagation procedure is determinant in the expression of *Banana streak virus* integrated into the genome of the FHIA 21 hybrid (*Musa* AAAB). Arch. Virol. 146(11): 2179-2190.

Fargette, D., Konate, G., Fauquet, C., Muller, E., Peterschmitt, M. and Thresh, J.M. 2006. Molecular ecology and emergence of tropical plant viruses. Annual Review of Phytopathology 44:235-260.

Fauquet, C.M., Mayo, M.A., Maniloff, J., Desselberger, U. and Ball, L.A. 2005. Virus Taxonomy: 8th report of the International Committee of the Taxonomy of Viruses. Elsevier Academic Press, Amsterdam.

Geering, A.D.W., McMichael, L.A., Dietzgen, R.G. and Thomas, J.E., 2000. Genetic diversity among *Banana streak virus* isolates from Australia. Phytopathology 90:921-927.

Geering, A. D. W., Olszewski, N. E., Harper, G., Lockhart, B. E. L., Hull R. and Thomas J. E. 2005. Banana contains a diverse array of endogenous badnaviruses. J. Gen. Virol. 86: 511-520.

Harper, G., Osuji, J.O., Heslop-Harisson and P., Hull, R. 1999a. Integration of *Banana streak badnavirus* into the *Musa* genome: molecular evidence. Virology 255:207-213.

Harper, G., Ganesh, D., Thottappilly, G. and Hull, R. 1999b. Detection of episomal Banana streak virus by IC-PCR. Journal of Virological Methods 79:1-8.

Lagoda, P.J.L., Noyer, J.L., Dambier, D., Baurens, F.C., Grapin, A. and Lanaud, C. 1998. Sequence tagged microsatellite site (STMS) markers in *Musacae*. Mol. Ecol. 7:657-666.

Le Provost G., Iskra-Caruana M.L., Acina I. and Teycheney, P.Y. 2006. Improved detection of episomal *Banana streak viruses* by multiplex immunocapture PCR. J. Virol. Meth. 137:7-13.

Lheureux, F., Carreel, F., Jenny, C., Lockhart, B.E.L. and Iskra-Caruana, M.L. 2003. Identification of genetic markers linked to Banana streak disease expression in inter-specific *Musa* hybrids. Theor. Appl. Genetic 106:594-598.

Lockhart, B.E.L. 1986. Purification and serology of a bacilliform virus associated with banana streak disease. Phytopathology 76:995-999.

Lockhart, B.E.L and Jones, D.R. 2000. Banana streak virus. p. 263-274. In: D.R. Jones (ed.), Diseases of Banana, Abaca and Enset, CABI Publishing, Wallingford, UK.

Ndowora, T., Dahal, G., LaFleur, D, Harper, G., Hull, R., Olzsewski, N. and Lockhart, BEL. 1999. Evidence that badnavirus infection in *Musa* can originate from integrated pararetroviral sequences. Virology 255:214-220.

Yang, I.C., Hafner, G.J., Revill, P.A., Dale, J.L., Harding, R.M. 2003. Sequence diversity of South Pacific isolates of *Taro bacilliform virus* and the development of a PCR-based diagnostic test. Arch. Virol. 148:1957-1968.

RESUME : Le génome des bananiers (*Musa* sp.) contient de nombreuses séquences virales EPRV (Endogenous pararetrovirus) appartenant au *Banana streak virus* (BSV), bien qu'aucun virus de plante n'ait d'étapes d'intégration dans son cycle. Certains EPRV provenant du bananier *M. balbisiana* sont infectieux car ils peuvent restituer des particules virales pathogènes en conditions de stress. La première partie de ce travail se focalise sur la biologie des EPRV. Nous avons tout d'abord analysé les caractéristiques moléculaires et génétiques des EPRV infectieux de l'espèce goldfinger du BSV (BSGFV) présents chez le bananier sauvage diploïde *M. balbisiana* cv. PKW. Nous avons ensuite identifié l'allèle infectieux de l'EPRV BSGFV, et abordé les mécanismes moléculaires de son activation par recombinaison homologue. L'évolution des séquences intégrées a été étudiée dans une deuxième partie. Une analyse phylogénétique à large échelle et une comparaison de l'évolution moléculaire des virus libres et EPRV chez trois espèces de bananiers nous ont permis de préciser l'origine phylogénétique des EPRV et de montrer que 27 évènements d'intégration indépendants se sont produits récemment dans les espèces hôtes. Nous avons ensuite étudié l'histoire évolutive de deux EPRV infectieux précédemment étudiés (BSGFV et BSV espèce Imové - BSImV) par l'analyse de leur polymorphisme de structure et de leur distribution au sein du genre *Musa*. Les résultats sont analysés en relation avec la phylogénie moléculaire des bananiers construite dans cette thèse. La probabilité d'intégration de chaque espèce de BSV est très faible, et à la différence d'autres pathosystèmes possédant des EPRV, il n'y a pas de colonisation des génomes hôtes par duplication des séquences virales une fois celles-ci intégrées. La forte diversité des EPRV chez le bananier s'explique plutôt par des évènements d'intégration indépendants de chacune des nombreuses espèces de virus libre.

Mots-clés : *Badnavirus*, Bananier (*Musa* sp.), *Banana streak virus* (BSV), EPRV infectieux, Séquences pararétrovirales intégrées (EPRV).

Evolution of plant pararetroviruses: the case of integrated sequences of *Banana streak virus* in the banana genome (*Musa* sp.)

ABSTRACT: The genome of banana plants (*Musa* sp.) harbours multiple endogenous pararetrovirus sequences (EPRVs) related to *Banana streak virus* (BSV), although no virus of plants needs integration for its replication. Some EPRVs of *M. balbisiana* are able to release infectious viral genomes under stress conditions resulting in viral infection of the plant. In the first part of our work, we focused on the biological characteristics of such EPRV. We described the molecular and genetic characteristics of an infectious EPRV of BSV Golfinger species (BSGFV) present in the wild diploid *M. balbisiana* cv. PKW. We identified the infectious allele of BSGFV EPRV, and proposed a model based on homologous recombination for its activation. In the second part, we studied evolutionary patterns of two EPRVs previously studied (BSGFV and Imové BSV species – BSImV). We first inferred large-scale phylogenies and compared the evolution rate and selective pressures acting on non integrated virus and EPRV found in three *Musa* species. We determined the phylogenetic origin of EPRV sequences found in these *Musa* species and estimated that at least 27 independent integration events occurred recently in the genome of the host species. Then, we studied the evolutionary history of the two infectious EPRV previously described (BSGFV and BSImV species) by analysing their distribution and polymorphism of structure among representative banana species, in relation to the phylogeny of *Musa* genus reconstructed in this study. The probability of integration of every species of BSV is very weak; and unlike other pathosystems harboring EPRVs, there is no colonization of host genomes by duplication of the viral sequences once integrated. The strong diversity of EPRV in the *Musa* genome could be rather explained by independent integrations from each of the numerous BSV species.

Key words: *Badnavirus*, Banana (*Musa* sp.), *Banana streak virus* (BSV), Endogenous pararetrovirus (EPRV), Infectious EPRVs.

Discipline : Biologie de l'Evolution et Ecologie
Unité où la thèse a été préparée : UMR BGPI - Biologie et Génétique des Interactions Plante-Parasite
CIRAD Département BIOS, TA A-54 /K Campus international de Baillarguet
34398 MONTPELLIER CEDEX 5, France